THE UNIVERSITY
of EDINBURGH

Self-directed learning in new and changing environments: Understanding human algorithms for exploration

Nicolas Collignon

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2019

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, either in whole or in part, in any previous application for a degree. Except where otherwise acknowledged, the work presented is entirely my own.

Nicolas Collignon

September 2019

Abstract

In order to act, plan, and achieve goals, people must learn about their environment and the outcome of possible actions. One reason for human successes in developing new theories and strategies when confronted with new problems is that people are not passive observers. Indeed, children ask informative questions and can adapt their strategies when inquiring about things they don't know. In this thesis, I aim to understand how people self-direct their learning across multiple tasks, when trying to achieve goals within their environment.

Chapter 2 presents experiments designed to better understand people's exploration and reward maximising strategies across a sequence of tasks. Through these experiments I examine how the environment, specifically the availability of information, and the prior knowledge of participants, affect their exploratory strategies. To study participant strategies I develop a general framework that considers both Bayesian approaches as well as a range of simpler heuristic strategies to approach the problem of goal-directed exploration (Chapter 3). This framework aims to explain the variation of participant strategies in terms of different underlying cognitive mechanisms that guide exploration. One of the benefits of a general framework is the ability to capture a diverse set of behaviours within a continuous parameter space. I focus on the problem of understanding the differences between participants by leveraging this shared psychological space. Specific families of strategies emerge from the behaviour of participants, highlighting

the importance of studying individual differences to better understand cognitive mechanisms (Chapter 4).

In Chapter 5, I analyse the experimental data from Wu *et al.* (2018) that considers similar phenomena concerning human exploration, with a specific focus on people's ability to generalise to guide their search. My analysis shows that our general framework offers a more compelling explanation for participant behaviour than the model they present, while again highlighting the importance of looking at individual differences. From these model based analysis we find that people are able to adapt to the structure of their environment, and are guided by local uncertainty rather than global uncertainty during exploration.

Finally, Chapter 6 looks at participants' behaviour when learning across a sequence of tasks when the underlying problem structures may change. How do people learn in a changing world? I show that the theory of inference by sampling can help explain distinct phenomena relating to the dynamics of learning across tasks. Our models are able to explain people's ability to progress across tasks when they share structural similarities, their ability to adapt to change, but also specific contexts where participants are continuously unable to realise the world has changed.

Acknowledgements

Quanto manca alla vetta ?

Tu sali e non pensarci!

Daniele Garozzo

I'd like to first thank my supervisor, Chris Lucas, for guiding me through and giving me the opportunity to take on this great journey of self-directed learning. I owe much to him for sharing his thorough scientific (and Bayesian) mind with me, his enthusiasm, intellect, many patient explanations, and all the time he's given me during these four years in Edinburgh. I feel very lucky to have been one of his very first PhD students. I'd also like to thank Alex Lascarides, Frank Keller and Mirella Lapata, for their feedback and encouragement during my yearly reviews, as well as my examiners, Neil Bramley and Adam Sanborn for helpful comments that helped improve the final version of this thesis.

I have spent countless hours in the direct vicinity of Pablo and will remember our late January Cogsci nights fondly. I must confess my admiration for his ability to deal with my incessant and silly questioning with humour and cool. Joseph Cronin was my greatest supporter during these four years – thank you for oddsing on it, and for being one of the kindest souls alive.

My academic journey has been greatly influenced by Esben Sørig who shared his passion for computational methods with me back at UCL, and remains a mentor and a great friend. I'm also indebted to Sofie and Arthur for passionate discussions about the brain, language and the mind in Acton Street. Without Carl and Amanda's BASc degree, and Dave Lagnado's unique approach to science, I very much doubt I would have had the opportunity to discover the field of computational cognitive science so quickly and I remain infinitely grateful.

The university fencing club offered much procrastination from research, but also continued to fuel my curiosity for people's ability to learn and adapt. I'd like to thank my coach Stan Stoodley for the many lessons, philosophical conversations and late night gossip after training, and to my teammates, particularly Mark, Etienne and Gino for their support.

Many parents pick up algebra again to help their children with homework - my mum thought it made sense to start a PhD to cheer me along, and continues to inspire me through and through. Thank you also to my dad for pushing me and always believing in me.

A PhD is a feat of endurance and these final words wouldn't be without the many people who encouraged me along the way. Merci to Ariathney - for being an endless source of exploration and curiosity since the start of my academic journey. Thanks to Tim Chen for his piano recital that one night in Bermondsey, Chris Nagle who helped me at the start of my research, Ramon for bringing Edinburgh back to life, Janie for the ciggie breaks, Martino for ending the coffee breaks short, Maria for sharing the last bit of the way, Ludovica for her perpetual positivity, Todor for shisha and RL, Kathryn O'B for the best craic, and Fiona for listening to me talk about models. A big thank you to Akash who was a true role model during my first two years in the Forum. I owe some cheers to the London crew who always made me feel at home down South. Thanks also to Vera for the Norwegian Wood breaks who punctuated research cycles, Svend for being a prime example of uncertainty directed exploration, Emily for dim sum in Hong Kong and her continual encouragements, as well as Will Sturgeon and Jonny Moens for helpful feedback on the final edits. Dr. Lachlan was a great companion during the Covid-19 lockdown and deserves a special note as he witnessed the final corrections.

I'm grateful to Professor Kando for hosting me at NII in Tokyo, and who expanded my perspective on academic research. Finally, I'd like to thank Veronique Mortensen for a formidable stay au Moulin in Coirac who helped me finish the writing of this thesis.

Contents

List of Tables

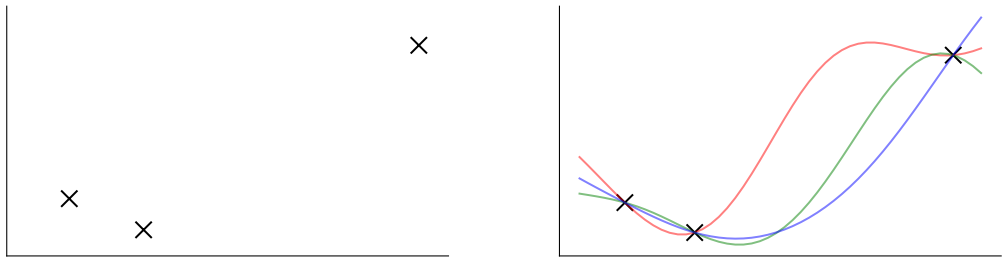
List of Figures

To my brothers, Martin, Vincent and Gabriel

Chapter 1

Introduction

Understanding the world emerges from the interaction of two processes. The first – empirical, or experimental – involves observing new phenomena and seeking information. The second – reflective, or theoretical – involves coming up with hypotheses to explain that which has been observed. The intricate interaction between both processes, experimental design and hypothesis formation, has been a subject of great interest in historical analyses of scientific discoveries. A common picture used to explain theory formation is the drawing of a line to fit data points. A successful theory will be one that explains the previous observations well, and perhaps more critically, one that is also able to predict future observations accurately.



(a) Example of three observations **(b)** Theories to explain past observations

Figure 1.1: Example of curve fitting to explain past observations. Which datapoint ($\mathbf{x}$) should one select next to learn the most about the unknown function?

Generating good theories about the world is one of the crucial abilities that enables people to solve a large and diverse set of problems across many domains. When confronted with new tasks, like playing a game, learning and applying a technical skill, or making friends in a new culture, people are able to rely on their previous experiences for guidance, without assuming that a new challenge is identical to a familiar one. One reason for human successes in developing new theories when confronted with new problems is that people are not passive observers. Throughout our lives, we learn by asking about our environment and interacting with it. We direct our attention and choose actions in ways that allow us to learn better and more efficiently than if we had to learn through observation only. Our agency helps us learn in two ways. First, each action we take is like a tiny experiment, allowing us to distinguish causal relationships from mere correlation (Sloman & Lagnado, 2005). Second, some events are more informative than others. Interacting with the world thus allows us to both inquire about the causal structure of the world, and to select actions that maximise information intake. Indeed, children ask informative questions and can adapt their strategies when inquiring about things they don't know (Ruggeri & Lombrozo, 2014). For example, a child might point to an object to know its name. They also play with new toys in ways that help them disambiguate uncertain causal relationships and

gather information (Schulz *et al.*, 2007; Schulz & Bonawitz, 2007; Cook *et al.*, 2011). The idea that people, like scientists, expand and refine their beliefs by performing intuitive experiments that are efficient (or even optimal) in providing us with information, e.g., about causal structure or the solution to a problem, has been particularly influential (see e.g. Gopnik *et al.*, 2004). This idea that humans perform intuitive experiments, maximising information gain, has been applied to understand causal learning (Bramley *et al.*, 2015), concept learning (Gureckis & Markant, 2009), question answering (Rothe *et al.*, 2016) and more (for an overview, see: Coenen *et al.*, 2017; Gureckis & Markant, 2012; Nelson, 2005; Schulz, 2012). In educational theory, it is also a widely known phenomenon that *active learning* has many benefits for the learner and leads to better outcomes than passive learners (Markant *et al.*, 2016a; Bruner, 1961; Kuhn *et al.*, 2000; Freeman *et al.*, 2014).

Learning, however, rarely happens for its own sake. While we are learning, we are often also trying to achieve particular goals, and there is a tension between choosing the actions that maximise the information we gain and those that are most likely to provide immediate rewards. A poker player might call a bet with a weak hand in order to learn about another player's propensity for bluffing, while a diner might choose an old favourite, even though visiting a new restaurant might have been better and more informative. This *explore-exploit* trade-off exists in a wide range of scenarios in science, medicine, industry, finance, policy making, and robotics, where we find a similar tension between either spending time to collect more data, or sticking to the information we currently have to take decisions that could probably be improved given better knowledge (Mehlhorn *et al.*, 2015; Hills *et al.*, 2015). Coming back to the example of fitting lines to data-points, this can correspond to the problem of finding the maximum of an unknown function within the fewest number of observations (see Figure 1.2). In this case, the theories generated to explain the data also aim to guide the selection of future rewarding

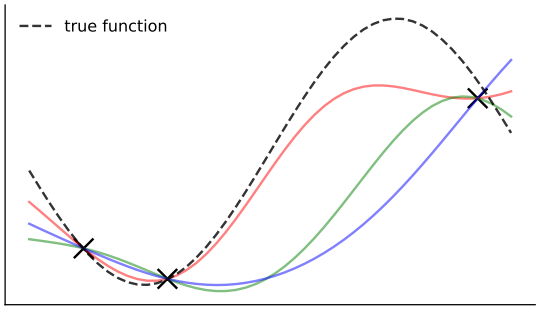


Figure 1.2: How do people direct their learning when trying to achieve goals (e.g. finding the maximum of an unknown function)?

actions. Should one select new actions because they are highly uncertain or actions that are close/similar to previous rewarding actions?

People's decision strategies when maximising rewards in the face of uncertainty have been studied extensively. Multi-armed bandit (MAB) problems, where the goal is to maximise reward sequentially choosing one of the N-arms of a bandit (a bandit can be visualised as a row of slot machines, also known as "one-armed bandits"), have been popular when studying how people balance exploration and exploitation. Studies with MABs have provided evidence that people behave following a rational strategy, with their behaviour accurately predicted by Bayesian learning models (Acuna & Schrater, 2009; Behrens *et al.*, 2007). Evidence suggests that people explore according to their uncertainty, consider the probability that an option offers the maximum reward, and are able to adapt to the volatility of their environment (Speekenbrink & Konstantinidis, 2015). Recent work shows that people's choices can be explained by a combination of uncertainty-directed and random exploration strategies (Gershman, 2018). Studies have also underlined consistent differences across individuals, relating different learning strategies to the psychological characteristics of participants (Steyvers *et al.*, 2009).

While MABs treat each arm as independent of the other, many real-world

problems require a learner to choose between actions that may share different degrees of similarity, indicating potentially similar outcomes. This is particularly true in problems with large decision spaces. In these types of problems, each action can be represented as a set of continuous and discrete features. Learning how features relate to rewards allows for an efficient representation of the environment, but also enables the learner to generalise to new events. For example, most shoppers can make a confident prediction of a fruit's ripeness based on its texture and colour. And the strength and angle at which a golf player hits the ball will describe how close it lands to a target at a driving range. These types of problems are formally referred to as *contextual* MABs (or CMABs), when the arms of a bandit have features that explain their reward distributions. In learning how contextual features predict rewards, an agent in a CMAB task is implicitly solving a *function learning* task, that is, using examples to discover relationships between variables.

A number of studies have explored how people learn functions, and Gaussian Process models have been proposed as a strong candidate to explain human behaviour in diverse function learning tasks, providing a unifying framework for rule-based and similarity-based models (Lucas *et al.*, 2015). These models have been useful in understanding how people exploit structured representations of the environment to direct their exploration in CMAB tasks (Schulz *et al.*, 2017b; Borji & Itti, 2013; Wu *et al.*, 2017). However, these experiments and the models associated with them have assumed that the basic structure of the underlying problem (i.e. which features relate to reward) is known in advance, or at least unchanging from one task to the next. However, the real world is rarely so obliging; it offers a multitude of different tasks that can change over time, some of which have structures that we don't know *a priori* (Gershman *et al.*, 2015).

Understanding how agents ought to take decisions that maximise cumulative reward has been the main focus of research in the field of reinforcement learning

(RL). Through its development, RL has contributed to our understanding of human decision making by providing a normative framework within which behaviour can be analyzed (Sutton & Barto, 1998; Niv, 2009). Recent advances have been met with notable successes in complex settings, like playing Atari video games at super-human performance levels (Mnih *et al.*, 2015) and defeating world champions in the game of Go (Silver *et al.*, 2016). One drawback of these algorithms is their need for massive amounts of data. Another is that, unlike humans, they have been unable to learn multiple tasks sequentially without forgetting previously acquired knowledge: a phenomenon referred to as "catastrophic forgetting" McCloskey & Cohen (1989). The challenge of continuously learning across multiple tasks remains largely unsolved and is a topic of considerable interest (e.g. see Kirkpatrick *et al.*, 2017; Wang *et al.*, 2016).

The broad aim of this thesis is to come to a closer understanding of the mechanisms that underpin human abilities to direct their learning to achieve goals in a complex and changing world. By studying people's strategies, we characterise human adaptive behaviour across sequences of tasks in order to understand how people detect change and similarities between tasks so as to avoid restarting their learning from scratch every time they are faced with a change of context.

In Chapter 2, I introduce the experimental paradigm that will form the empirical basis of this thesis. Through four experiments, I study the learning strategies of people when faced with a sequence of new but related tasks. I examine how the environment, specifically the availability of information, and the prior knowledge of participants, affect their exploratory strategies. Empirically, I find that people display biases and patterns of behaviour that deviate largely from the qualitative predictions of Bayesian rational models. Rather than constructing a model of their environment to select actions that maximise information or rewards, many participants relied on the use of simpler heuristic strategies. In Chapter 3, I develop a general modelling framework to better understand the strategies and

representations of people during learning and with the aim of capturing a diversity of behaviours under a unique parameter space. The empirical analysis of Chapter 2 pointed out significant differences in the strategies used by participants. In Chapter 4, I leverage the framework of our general model to study patterns in variation of participant strategies. Four families of strategies emerge from the behaviour of participants, recurring across the different experimental conditions. While the large majority of participants relied on previous observations to predict the outcome of unknown actions, only a minority of participants seemed to rely on uncertainty to guide their search. Instead, we find that novelty and local search were prominent factors behind the decision strategies of many participant.

In Chapter 5 I look more specifically at people's ability to generalise from previous observations. I analyse the experimental data from Wu *et al.* (2018) that presents different reward structures in an experimental paradigm similar to the one we introduced in Chapter 2. An initial empirical analysis shows patterns in individual differences similar to the ones discovered in Chapter 2. I then show that our general framework offers a more compelling explanation for participant behaviour than the model they present, and I offer a different interpretation of the psychological claims one can draw from the modelling results. First, I show that people are able to generalise adaptively to the structure of their environment. Second, and contrary to the picture put forward by e.g. Schulz & Gershman (2019), I show that (global) uncertainty-directed search is rarely the main driver for the exploratory strategies of people. Instead, I argue that people are driven by expected rewards, local uncertainty and novelty. Overall, I demonstrate the importance of taking individual differences into account when studying human behaviour.

Finally, Chapter 6 presents empirical data for people's ability to learn across sequences of tasks when the underlying problem structures may change. Through model simulations, I show that inference by sampling can help explain distinct

phenomena relating to the dynamics of learning across tasks, namely people's ability to progress across tasks when they share structural similarities, their ability to adapt to change, but also the specific contexts in which participants are continuously unable to realise the world has changed.

Chapter 2

Epistemic drive and memory manipulations in explore-exploit problems

2.1 Introduction

In the previous chapter we introduced the general problem of understanding how people direct their learning in order to achieve goals in sequences of tasks. In this chapter, we study how people learn to select actions that are most rewarding when faced with a sequence of novel but potentially related tasks, and more specifically how people might make use of contextual features to guide their learning. We present the results of four experiments designed to better understand people's exploration and reward maximising strategies across a sequence of tasks. Do those strategies evolve over time, as they encounter related tasks? Can people transfer structural knowledge and improve their performance by leveraging similarities

between tasks? What is the relationship between people's search strategies, their ability to learn and generalise from observations, and how well they do?

When encountering new situations, people are often faced with a tension between gathering more information to improve the quality of their future decisions, or choosing actions that are known to be rewarding (Hills *et al.*, 2015). A doctor might, for example, want to run more tests to have a better diagnosis for their patient, or instead choose to give them the treatment they believe will best relieve them from their symptoms. Multi-armed Bandits (MAB) have been a popular experimental paradigm to better understand human decision strategies when dealing with the explore-exploit trade-off (Cohen *et al.*, 2007). The metaphor of the multi-armed bandit comes from the rows of slot machines in casinos, where a gambler has to sequentially choose the arm of a bandit to maximise their gains. In the same way, participants in these experiments have to choose between different possible actions (e.g. the arms of a bandit) so as to maximise their rewards. The learner is thus faced with a trade-off between choosing rewarding actions, or actions that will provide more information about the reward distribution of the different options. In this case, the arms of the bandit yield stochastic rewards and are independent of one another.

In the real world, an essential part of solving problems lies in discovering their underlying structure. Actions can often be represented as a set of features (e.g. how hard to push a button, or its position on a screen), and the outcome of an action might give information about other similar actions. For example, two structurally similar chemical compounds might have similar properties, or animals of similar size might produce sounds of a similar pitch. These properties are present in Contextual MABs (CMAB), where observable features provide information about the arms' reward distributions. Learning how features relate to rewards allows for an efficient representation of the environment, and enables the learner to generalise to new events.

The task of the learner can thus be separated into two distinct types of generalisation when learning across a sequence of tasks. On the one hand, the learner must extrapolate from the limited observations within their current context to infer the latent structure of their environment. By learning the relationship between action features and their outcomes, one can predict the outcome of new and unobserved actions. We call this type of learning *within-task* generalisation. This process has been widely studied in the context of category learning, where one has to predict the class of an unknown object or entity given their descriptive features. Similarly, the domain of function learning has focused on the human ability to predict for continuous relationships (as opposed to categorical ones) to better understand the human biases that guide learning. Learning the functional form of relationships allows us to predict e.g. the amount of pressure one needs to exert on the acceleration pedal to get to a specific speed, or the ripeness of fruit based on its colour and softness. The outcome of this research has shown that while people exhibit a strong bias toward linear relationships (Brehmer, 1976; Kalish *et al.*, 2004; Busemeyer *et al.*, 1997), they are able to learn a wide range of relationships that allow them to generalise efficiently about unobserved data (Lucas *et al.*, 2015; Wilson *et al.*, 2015; León-Villagrá *et al.*, 2018).

The second task of the learner involves generalising from one task to the next, or *across-task* generalisation. In an ever changing world, the same situation is never encountered twice, yet we are able to leverage previous experiences to make decisions and achieve goals. This can be done when contexts share structural similarities. By constructing representations that capture abstract similarities (Gershman *et al.*, 2010a) or aspects of a task that are subject to change (Wilson & Niv, 2011), the learner is able to deploy efficient generalisations across a wide range of scenarios.

Previous studies of human behaviour in CMAB problems have shown that people are able to generalise from their observations when faced with a large number of

options, and make use of uncertainty to direct their search (Schulz *et al.*, 2017b; Wu *et al.*, 2018; Borji & Itti, 2013). These experiments have assumed the basic structure of the underlying problems to be static, or known in advance, thus only focusing on *within-task* generalisation. When confronted with unknown task structures, Teodorescu & Erev (2014) showed that people are able to adaptively learn purely exploratory or purely exploitation-oriented policies. However, in their experiment there was no systematic relationship between an option's features and its reward, thus leaving out *within-task* generalisation. Here, we study how people select actions that guide both processes.

Unlike a CMAB-type task, the tasks we presented to participants were deterministic, meaning that re-selecting an option would always yield the same reward. This was done to ensure a clear distinction between exploration and exploitation in participant decisions. In our experiments, each action had a set of features (e.g. the brightness of a button or its location in space). The features associated with an action were predictive of its associated reward. To examine people's ability to generalise we presented them with tasks that contained a large number of choices and a relatively limited number of actions. In this case, generalising over previous observations is necessary for optimal performance. We chose a simple structure to ensure it would be possible for participants to learn and exploit it when maximising rewards.

This chapter presents the empirical results from four experiments. Our first two experiments focus on sequential tasks where participants had no prior information about the underlying reward structure, and where a combination of exploration – to discover task structure and discover optima – and exploitation is necessary to do well. The next two experiments provided participants with training about the reward structures before the task itself to understand the effect of prior knowledge on participant strategies.

We initially predicted participants would be well accounted by Bayesian models in our tasks. In general, these models first entail constructing a model of the world and updating this model in light of new evidence. Second, they consist of policies on how to select future actions based on those beliefs – namely the uncertainty one holds about the value of unobserved actions, and their expected outcomes. More specifically, and more qualitatively speaking, we predicted participant would select globally informative actions guided by their measure of uncertainty in order to learn the underlying task structure. We also predicted participant would be able to trade-off between exploration and exploitation, and favour rewarding actions after a preliminary phase of exploration. Finally, we expected participants would be able to improve their performance across similar tasks by re-using the structure learned in previous tasks.

To foreshadow, we find that, across multiple experiments, some participants selected actions that resolved uncertainty about the underlying structure of the task, and traded off between exploration and exploitation in order to maximise reward. These participants were able to transfer knowledge across tasks and gradually improved their performance. On the other hand, a significant number of participants engaged in purely exploratory behaviour, consistently preferring to choose novel actions despite them being relatively unrewarding. Our results highlight the importance of studying individual differences to better identify the multiple factors that influence human behaviour, and of accommodating these differences in models of learning and exploration.

2.2 Experiment 1

Across our four experiments, participants were given a sequence of grids composed of 9-by-9 arrays of tiles (see Figure 2.1), with each tile corresponding to a possible

action. In this chapter, we limit our analysis to the first three grids presented to participants (out of nine), as the latent task structure changed after that point. We study the behaviour of participants when the underlying structure may change in Chapter 6. The grids studied here shared a similar underlying task structure: they had the same kind of relationship between features and rewards, but details of those relationships varied. In our experiment an action consists of selecting an individual tile, which has two features: its horizontal (x), and vertical position (y). Participants had to select tiles to maximise their cumulative rewards over 20 choices in each grid. The task presents a classical explore-exploit trade-off: Succeeding requires carefully balancing between choosing new tiles to learn about the underlying reward structure or re-selecting tiles that were observed to be rewarding. In Experiment 1, participants received no prior knowledge about the reward structure of the tasks, nor about whether the tasks were related to one another in any way.

The aim of our first experiment was to understand how people select actions to learn about the structure of their environment, and achieve goals within it – and do so across a sequence of potentially related tasks. Motivated by people's observed ability to extrapolate from sparse data and learn functional relationships (Lucas *et al.*, 2015; Wilson *et al.*, 2015; León-Villagrá *et al.*, 2018), and by the successes of Bayesian models in active search tasks (Schulz *et al.*, 2017b; Wu *et al.*, 2018; Borji & Itti, 2013), we predicted participants would be able to generalise from previous observations and improve by using their growing knowledge of the underlying task structure to select better actions. We measure this by looking at participants' ability to select more rewarding tiles as they collected more information within the same grid, and whether they demonstrated confidence in their knowledge by repeatedly selecting (i.e., exploiting) optimal actions. Given people's ability at taking advantage of the structure inherent in real-world tasks when learning and trying to achieve goals (Gershman *et al.*,

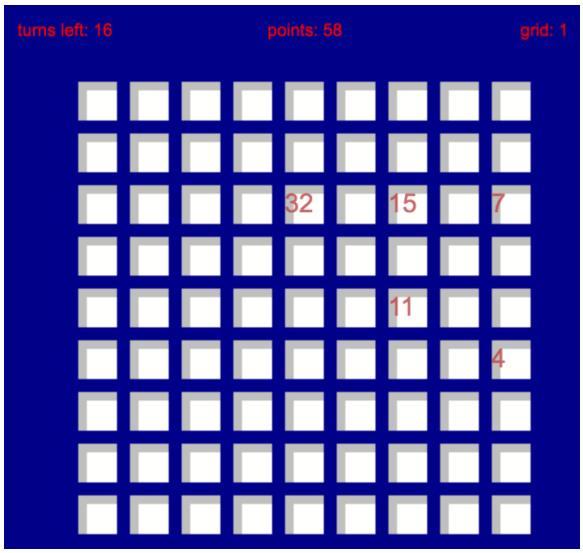


Figure 2.1: Grid presented to participants after 5 observations. Note that in Experiment 1, the rewards disappear shortly after a tile has been selected (1.5 seconds).

2010a; Wilson & Niv, 2011), we predicted participants would be able to re-use knowledge across grids, since they shared the same structure, and would improve their performance from one grid to the next.

We also studied the (Euclidian) distance between participants' selections throughout the task to better understand their behaviour. Distance between selections is a useful marker of different exploration strategies. For example, participants who seek to reduce uncertainty about the task structure are likely to select tiles that are far apart from each other, as these tend to yield more information about the broad shape of the reward function, in addition to having more uncertain rewards themselves. We call these selections *globally informative* actions. In contrast, participants might sample tiles adjacent to their previous observations, e.g., because they believe they are close to a maximum or because they want to observe local gradients. We call this kind of selection *local search*.

2.2.1 Methods

We recruited 79 participants using Amazon's Mechanical Turk service. They received \$0.75-\$1. Participants were told their rewards would be doubled if their final scores were in the top 10 percent. Following the instructions given to participants, we excluded participants whose performance was worse than chance ($n = 3$). We also excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 5$), as it showed a lack of engagement with the task.

The three grids analysed here used a reward structure where one location (x_m, y_m) was sampled uniformly at random in each grid, and the grids' maximum rewards m were sampled independently from $\mathcal{N}(\mu = 200, \sigma^2 = 50)$. The reward r for a given tile location (x, y) was exponentially decreasing with its Euclidean distance d from that maximum-reward tile rounded to the nearest integer:

$$r(x, y) = C \cdot e^{-k \cdot d((x,y),(x_m,y_m))}$$

We chose an exponential relationship between features and rewards to ensure there would be a clear advantage for participants who discovered the maximum-reward tile. We chose a constant ($k = 0.4$) that led to large differences between the maximum and its closest neighbors while making it unlikely that any tiles would have rewards of zero or one. We used a random maximum reward in order to make it difficult for participants to know they had found the most rewarding tile without knowing the reward structure of the task.

When a tile was selected, the reward was displayed on the tile for 1.5 seconds and added to the cumulative score on the current grid. Participants could re-select tiles they had previously chosen. Participants were only told there may be

patterns underlying the distribution of rewards, but were not given any explicit information. There were no cues or markers to indicate previously selected tiles throughout a grid.

2.2.2 Results

For this and all subsequent experiments, we report the normalised scores (between 0 and 1), by dividing each reward by the maximum possible reward in that grid. We were first interested in seeing whether participants were able to recognise similarities between tasks. We use a general linear model (GLM), with the reward as outcome variable. The turn and grid index were used as predictor variables. Both the turn ($b = 0.02, se = 0.001, p < 0.001$) and the grid ($b = 0.05, se = 0.005, p < 0.001$) were significant factors. Following our hypothesis, participants selected better tiles over time, suggesting that they were able to exploit the underlying reward structure. Participants also improved their performance across grids, suggesting they were able to transfer structural knowledge across tasks (see Figure 2.7).

As a simple measure of a participant's propensity to explore, we used the proportion of actions that selected a previously unseen tile ("exploration") versus re-selecting a previously seen tile ("exploitation"). This distinction is more natural in our tasks than in a traditional stochastic bandit task, as in the latter it can be informative to re-select previously-seen tiles to learn about their reward distributions. There were substantial behavioural differences in how people traded off between exploration and exploitation among participants, and in the cumulative rewards they collected ($M = 0.49, SD = 0.30$) (see Figure 2.2).

Twenty-two participants (31 percent) rarely re-selected tiles more than twice in the majority of grids. We call these participants *full explore* (FE) participants.

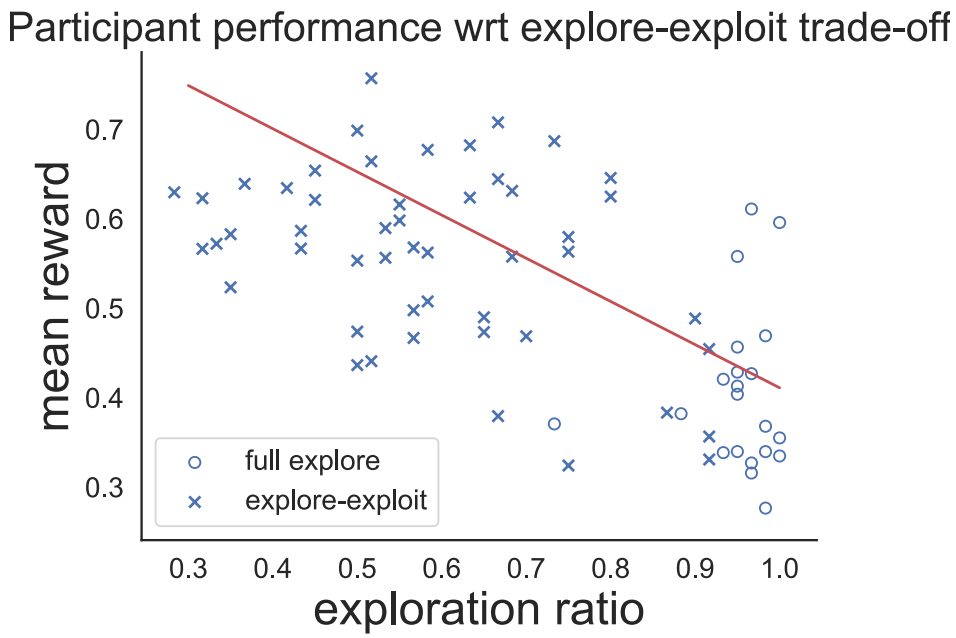


Figure 2.2: Each point represents a participant. The y-axis shows the average reward of a participant across all three grids. The x-axis is the proportion of novel selections across all three grids. A value of 1 would mean only selecting new tiles, 0 only selecting the previously-selected tiles.

We chose this criterion to account for possible accidental re-selections, and the possibility that participants would change strategy in one of the three grids. This followed the assumption that if a participant never re-selected a tile in two grids but balanced exploration and exploitation in the third, they would still be best described as a *full-explore* participant.

We call the other participants (n=49), that traded off exploration and exploitation, *Explore-Exploit* (EE) participants. We show examples of FE participants in Figure 2.3 and 2.4, and of EE participants in Figure 2.5 and 2.6.

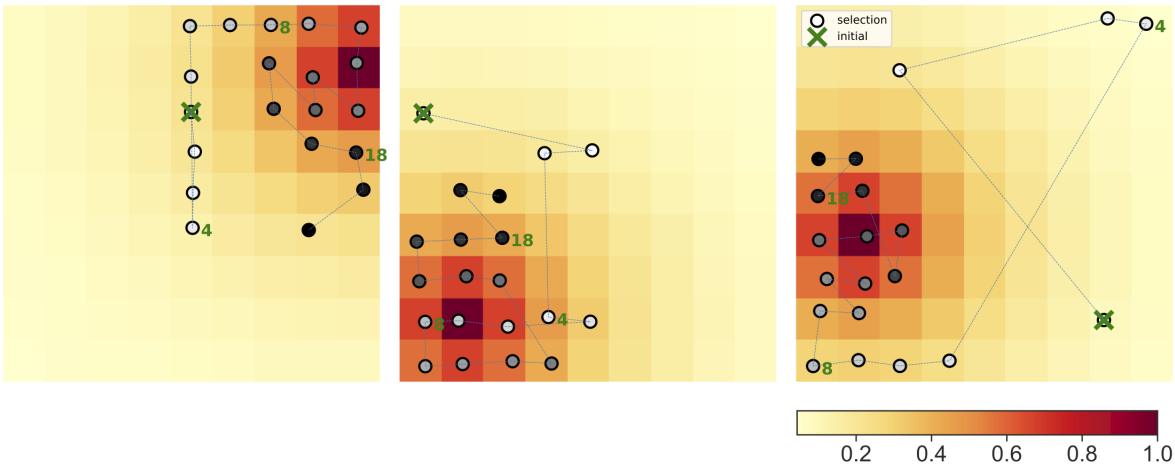


Figure 2.3: Selections of a *Full Explore* participants across all three grids. The green cross marks the initial observation. Markers indicate observations, and a darker shade means an observation was later in the trial. Numbers indicate the index of the closest observation marker. The colour of a tile indicates its associated reward value (the darkest is the maximum).

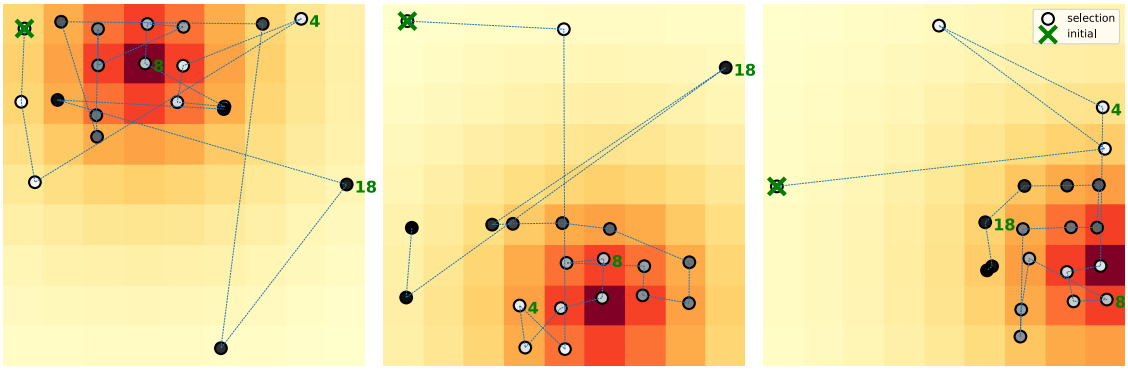


Figure 2.4: Selections of a *Full Explore* participants across all three grids.

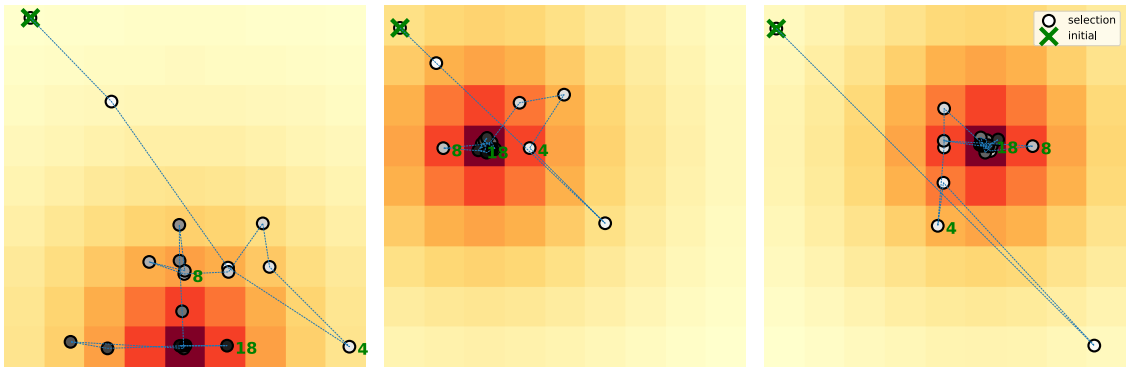


Figure 2.5: Selections of an *Explore-Exploit* participants across all three grids.

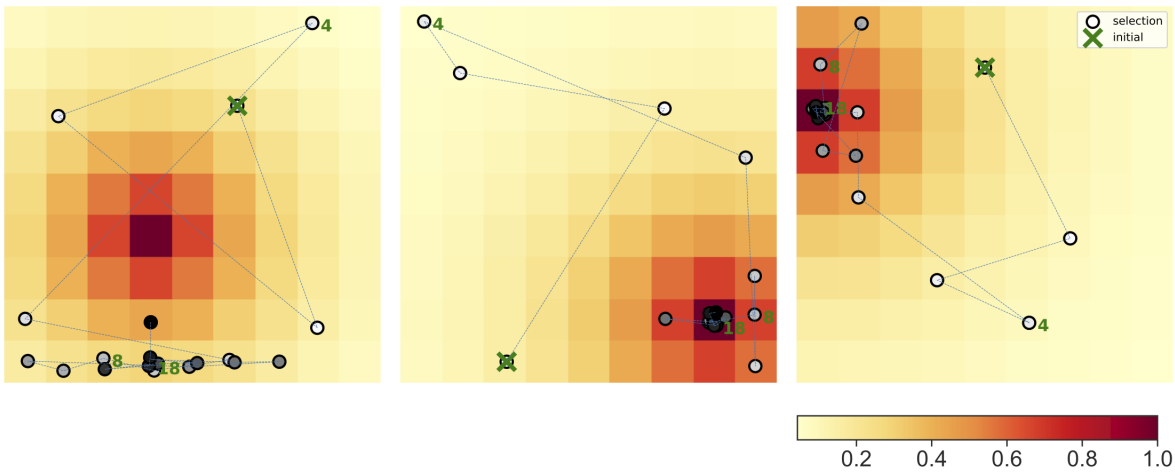


Figure 2.6: Selections of an *Explore-Exploit* participants across all three grids.

EE participants improved across tasks ($b = 0.07, se = 0.006, p < 0.001$) (see Figure 2.7), supporting our hypothesis that participants who used the underlying task structure to direct their search were able to re-use what they had learned to a new task.

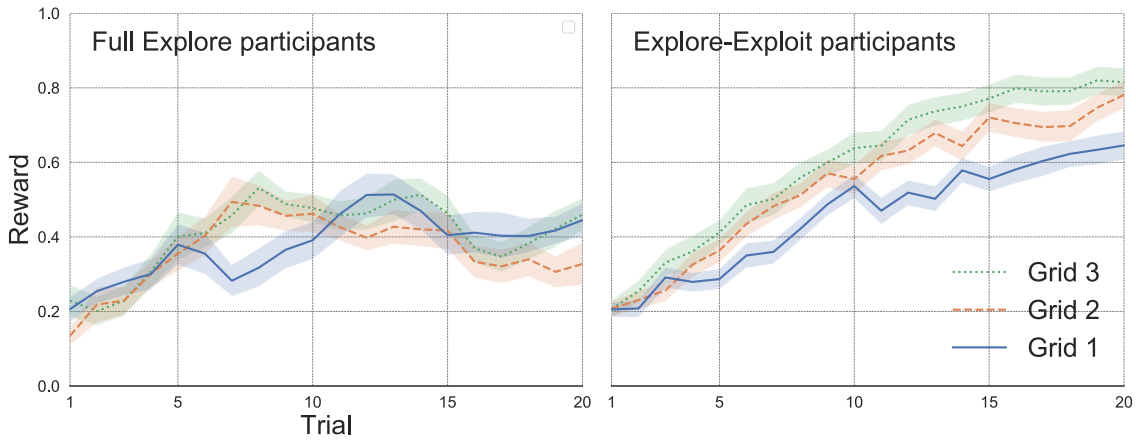


Figure 2.7: Performance of FE participants (n=22) and EE participants (n=49) in Experiment 1 across all three grids. Error bars in this and all subsequent plots reflect standard errors of the mean.

Across all participants, the proportion of exploratory selections correlated negatively with score ($r(140) = -0.71, p < 0.001$), and FE participants earned lower scores than EE participants ($t(69) = 5.77, p < 0.001, d = 0.15$). Their average scores barely improved from one grid to the next (Figure 3; $b = 0.02, se = 0.008, p = 0.06$).

We used a logistic regression model to evaluate participants' ability to find the maximum across grids. More participants found the maximum as they went on with the grids, hinting that they were better at utilising the underlying task structure ($b = 0.64, se = 0.11, p < 0.001$). Whether participants were engaging in *full exploratory* or *explore-exploit* strategies did not predict if they found the maximum in the tasks ($b < 0.001$). Participants were significantly better than

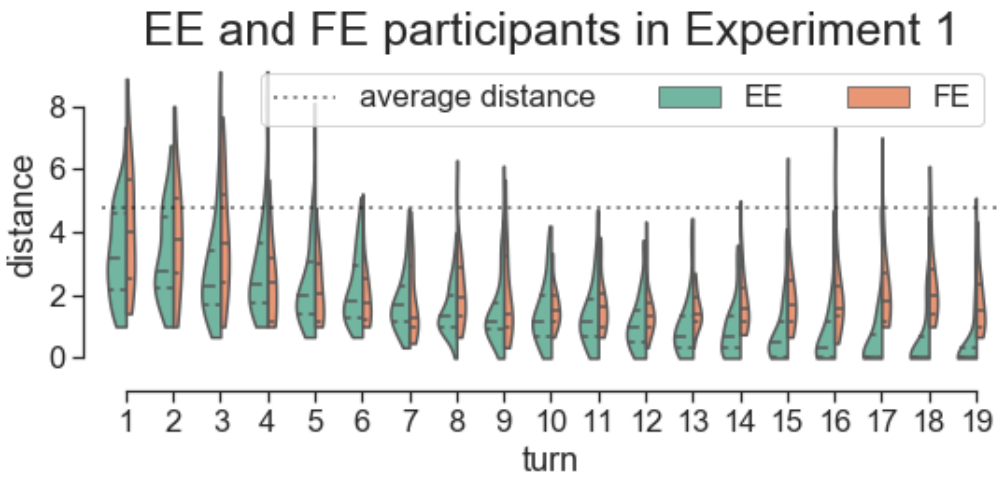


Figure 2.8: Average distance between selections of EE and FE participants in Experiment 1 at each turn. Distribution over distances are presented with quantiles and kernel density estimations. We use Euclidian distance between selections, with 0 counting for a re-selection of the previous click. The dotted line represents the average distance between all tiles in a grid. The shape of the distribution is drawn using a (normal) Gaussian Kernel Density Estimate.

chance at finding the maxima (0.65 of grids, vs. upper bound chance proportion of 0.25; $\chi^2(1, N = 1174) = 188.1, p < 0.001$).

Participants had overall a strong 'local bias' in their sampling, where they choose tiles close to their last choice more often than chance given the distribution of inter-tile distances ($t(151) = -50.8, p < 0.001, d = -2.34$) (see Figure 2.8). This suggests that participants engaged in local search strategies, rather than globally informative actions. Both EE and FE groups showed this bias, with adjacent tiles selected in 49% of FE participants' exploratory choices ($SD = 0.17$) and 39% for EE participants ($SD = 0.17$).

2.2.3 Interim discussion

Experiment 1 showed that some participants were able to learn the underlying task structure when it was new and traded off between exploration and exploitation to maximise their rewards. These participants transferred knowledge across tasks that shared similarities in their underlying structure. However, a large proportion of participants had a strong tendency to explore in circumstances where exploitation would have yielded much higher scores, preferring unobserved tiles over known tiles with a high reward value. Why did so many participants adopt such an extreme exploratory policy? One possibility is that they were motivated to learn more about the reward structure, or ensure they had found the maximum possible reward, in line with the inherent curiosity bias observed in people (Kidd & Hayden, 2015; Gottlieb *et al.*, 2013). We also considered this may be owing to a misunderstanding of the instructions or a lack of incentive, though FE participants presented some evidence for learning the underlying structure, even if this was not reflected in their score.

We also observed a locality bias in participants' choices. This may have been due to the memory demands of the task. Generalising from unavailable observations might be particularly taxing, and could have led participants to adopt policies that alleviated the complexity of the task. For example, if participants tracked local gradients in rewards and followed increasing rewards, this would only require tracking 2-3 past observations while being less demanding than computing a surrogate model over the general task structure. This would be consistent with the local search strategies exhibited in other domains such as causal learning (Bramley *et al.*, 2015) and category learning (Markant *et al.*, 2016b), and the idea that people adapt their high-level strategies to make the most of limited resources (Lieder *et al.*, 2014). For FE participants, the local bias during exploration could reflect a systematic and memory-efficient policy for exhaustively searching a subset of the tiles for a maximum. We also consider the possibility that

participants were "lazily" biased toward nearby tiles as it implied less distance between selections.

In Experiment 2, we presented participants with the same task structure as in Experiment 1, but with changes designed to understand and potentially reduce their strong tendency to explore new tiles. These included persistent indicators of explored tiles' rewards, checks of participants' understanding of the instructions and a different incentive structure.

2.3 Experiment 2

In this experiment, the reward associated with a given tile is displayed continuously once it has been observed. We hypothesised that with observations remaining visible, the overall reward pattern would be more evident. We predicted that participants would be able to make more globally informative actions (implying that exploratory selections would be more distant from each other). Because of the underlying structure being more evident, we also assumed fewer participants would engage in *full exploratory* behaviour.

2.3.1 Methods

We recruited 72 participants using Amazon's Mechanical Turk service identically to Experiment 1. Participants all received a base payment of \$0.75. The reward scheme differed from that in Experiment 1: rather than granting bonuses to the top 10 percent, we gave all participants a bonus proportional to their cumulative score, up a maximum of \$0.75. We excluded participants who failed to select more than 2 different tiles on the majority of grids ($n = 4$). In Experiment 2

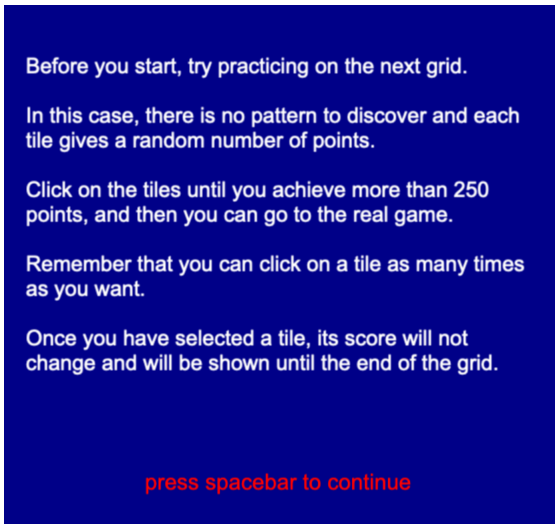


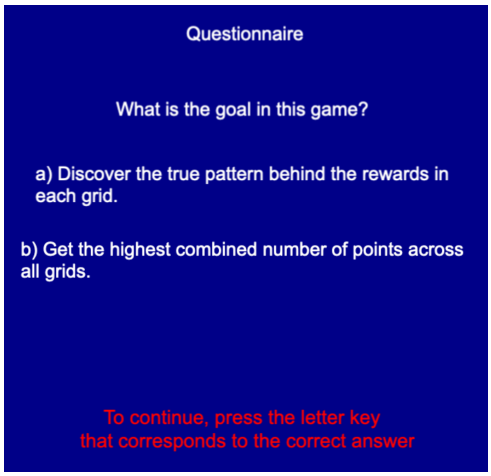
Figure 2.9: Experiment 2: Instructions prior to practice grid. Participants were told explicitly they could re-select, and were informed the scores would remain observable.

when a tile is selected by a participant the reward is continuously displayed on the tile and is added to the current cumulative score on the current grid.

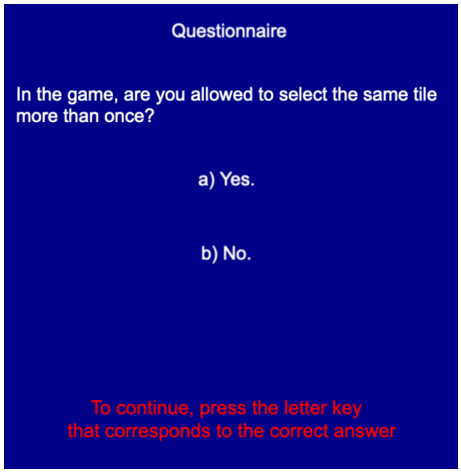
We added explicit instructions to tell participants they could re-select tiles (see Figure 2.9), and added a pre-task questionnaire to make sure participants understood these instructions. The questionnaire also required participants to understand their goal was to maximise reward (as opposed to discovering the underlying pattern, or finding the maximum). Participants were not allowed to proceed with the task until they answered all questions correctly (see Figure 2.10).

2.3.2 Results

We predicted that participants would be less prone to *full exploratory* behaviour (FE) because of the observable rewards. Contrary to our prediction, a significantly larger proportion of participants showed FE behaviour in Experiment 2 than in Experiment 1 (.47, $n = 32$ vs. .31, $n = 22$; $\chi^2(1, N = 139) = 18.6, p < 0.001$). As



(a) Questionnaire part 1.



(b) Questionnaire part 2.

Figure 2.10: Experiment 2: Pre-task questionnaire. Participants were not allowed to continue before having selected the right answer.

in Experiment 1, the proportion of exploratory selections correlated negatively with performance ($r(134) = -0.75, p < 0.0001$). In Experiment 2, *explore-exploit* (EE) participants again performed significantly better than FE participants ($t(66) = 9.31, p < 0.0001, d = 0.23$). Conducting a similar GLM analysis as in Experiment 1, we found that EE participants improved significantly across tasks ($b = 0.04, se = 0.007, p < 0.0001$), whereas FE participants did not ($b = 0.01, se = 0.006, p = 0.14$).

To understand the effect of having observations available throughout the task, we compare the performance of EE participants in Experiment 2 (EE_2, n=36) to the performance of EE participants in Experiment 1 (EE_1, n=49). Overall, EE_2 participants (M=0.58) did slightly better than EE_1 participants (M=0.56) ($b = 0.04, se = 0.008, p < 0.001$).

This was most pronounced in the first grid ($t(84) = 2.18, p = 0.03, d = 0.08$). We conjecture that EE_2 participants learned the reward pattern faster, and EE_1 participants caught up in subsequent grids. This supports the hypothesis that

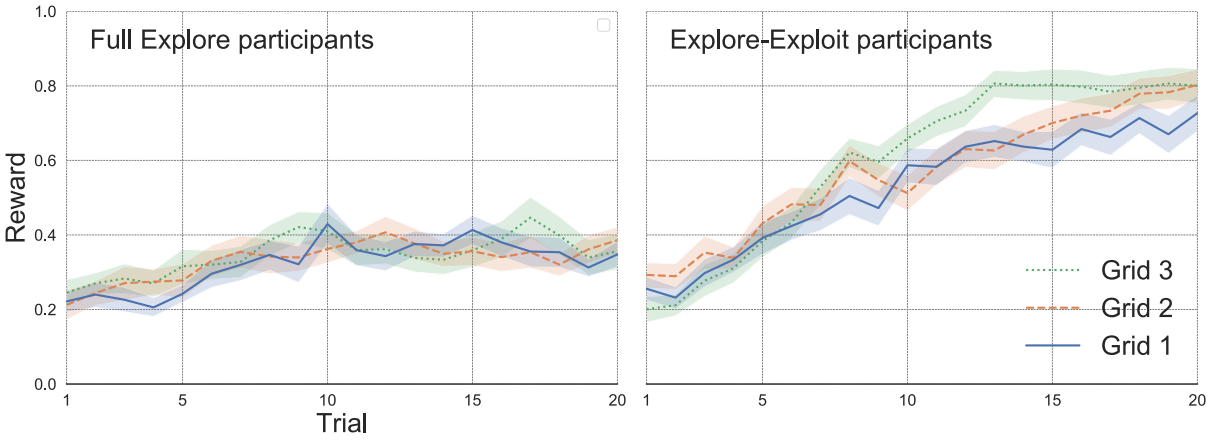


Figure 2.11: Performance of FE participants ($n = 49.$) vs EE participants ($n = 36$).

visible observations allowed participants to generalise better, by supporting more global strategies. To test this idea, we looked at the inter-selection distances between the initial selections of participants. We analyse the first 5 selections assuming that the most informative actions would be taken in the early stages. As predicted, the choices of EE_2 participants were more global, with greater distances than EE_1 participants' choices ($t(84) = -2.25, p = 0.03, d = 0.66$) (see Figure 2.12).

2.3.3 Interim discussion

Experiment 1 showed that a large number of participants engaged in full exploratory behaviour. In Experiment 2, we looked at the effect of leaving observations visible once they had been selected, under the assumption that it would make generalisation easier for participants, and thus make them less prone to over-explore. Experiment 2 showed that some participants were indeed able to leverage visible observations to conduct more global exploration, which led to better overall performances. However, the observable rewards also seemed to add a further incentive for many participants to exclusively choose novel actions, rather

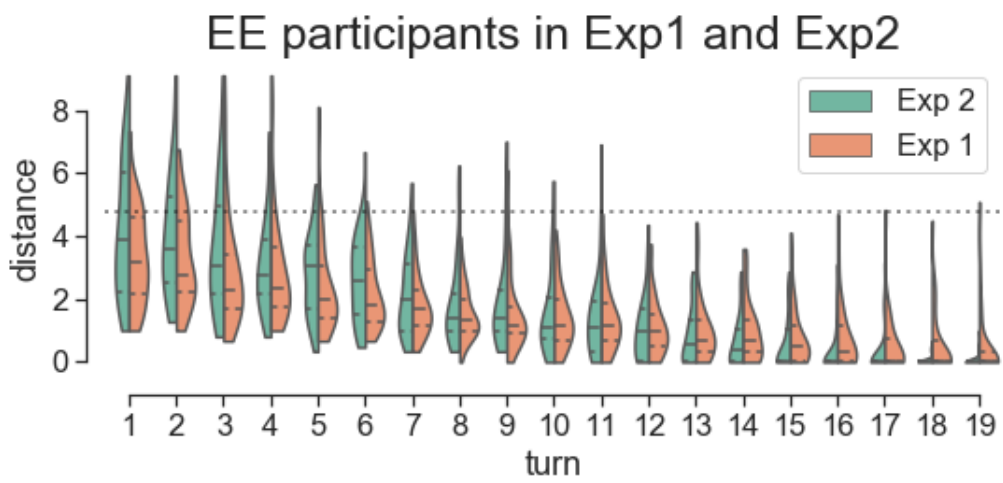


Figure 2.12: Comparison of distances between selections of EE participants in Experiment 1 and Experiment 2 (see Figure 2.8 for details). EE participants in Experiment 2 selected more "global" actions (longer distances between selections) during their first actions.

than maximising rewards. Why did more participants engage in full exploratory behaviour in Experiment 2? We conjecture that participants might have been more motivated to observe rewards for new tiles when these remained visible, because the overall pattern – and the possibility of better understanding it – might have been more salient to them. In Experiment 3, we sought to better understand why some participants might want to select new tiles almost exclusively, rather than occasionally exploiting what they had learned to earn greater rewards. After Experiment 1, we suggested that this exploratory behaviour might be due to an intrinsic epistemic drive in participants. We hypothesise this behaviour will only occur for new tasks when participants have no prior knowledge about the underlying reward structure of the tasks, since new observations would not be very informative if participants are already familiar with the underlying reward structure.

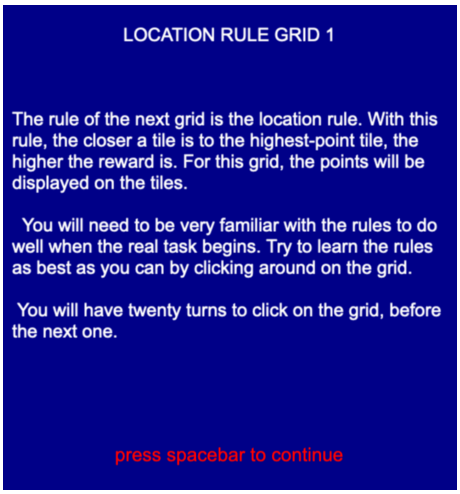
2.4 Experiment 3

We designed Experiment 3 to control explicitly for the potential epistemic drive of FE participants by familiarising them with the underlying reward structures prior to the task. By making the structure clear to participants prior to the tasks, our primary prediction for Experiment 3 was that fewer participants would engage in FE behaviour. We presumed the intrinsic motivation of observing new observations would be attenuated when participants did not gain new information about the task from those observations.

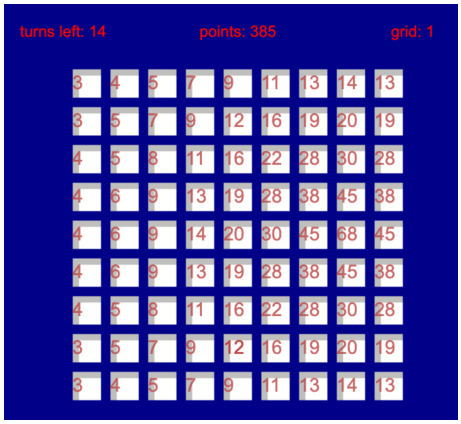
We also hypothesised there would be weaker or no progress across grids since participants would already be familiar with the reward structure when they engage with the first grid. Because of the training, we predicted participants would be more efficient at finding and re-selecting tiles with high values, and would thus perform better overall than in Experiment 1 and 2. Experiment 3 was set up identically to Experiment 2. Participants were told about the underlying pattern and given three practice grids so they could learn the reward structure prior to the task.

2.4.1 Methods

We recruited 43 participants using Amazon's Mechanical Turk service, identically to Experiment 2, with the following changes: Participants were only recruited for three grids rather than nine, following the same reward pattern discussed in Experiment 1 and Experiment 2. We report on the 6 later grids from Experiment 1 and 2 in Chapter 6. Because of the shorter duration, participants were paid a base reward of \$0.2. We used a proportionally larger bonus of \$0.6 under the logic that this would further reduce the relative effects of epistemic drive. Apart from the training grids presented prior to the task, instructions were identical to



(a) Explanation of the Location rule given to participants prior to the training grids.



(b) First practice grid given to participants to familiarise themselves with the Location rule.

Figure 2.13: Location rule instructions. The two next training grids do not show the rewards for all tiles, but participants are told explicitly to select tiles in order to learn the reward pattern.

Experiment 2. During the training, participants were told that each grid had one maximum tile, and the closer a tile is to the maximum the higher the reward (see Figure 2.13). The first training grid had all rewards displayed and participants were instructed to familiarise themselves with the nature of the task. The next two grids were similar to the grids in the actual task (i.e. only observed tiles display reward values) but participants were encouraged to learn the pattern as well as they could. Throughout the task, instructions regarding reward maximisation and the possibility of reselecting tiles were also displayed. We excluded one participant who failed to select more than two different tiles on the majority of grids and one participant who reported not following the instructions upon completing the experiment.

2.4.2 Results

Surprisingly, 37 percent (15 out of 41) of participants engaged in *Full Exploration* (FE) in Experiment 3. We used the same criterion as in Experiment 1 and 2. The proportion of FE participants in Experiment 3 was significantly less than the 47 percent we observed in Experiment 2 ($\chi^2(1, N = 109) = 8.82, p = 0.003$), but was nonetheless a higher proportion than anticipated.

As expected, EE participants in Experiment 3 did not improve significantly across grids, since they had been trained extensively on the rule before the assessed task started ($b = -0.01, se = 0.008, p = 0.112$). The average performance of EE participants was significantly better than EE participants in Experiment 2 ($t(61) = 2.29, p = 0.03, d = 0.07$) and EE participants in Experiment 1 ($t(74) = 3.11, p = 0.003, d = 0.09$), suggesting that participants were able to learn the rule during the training and relied on this knowledge during the tasks.

To understand the effect of prior knowledge on participants' exploratory patterns, we compared how EE participants in Experiment 3 (EE_3) explored compared to EE participants in Experiment 2 (EE_2). EE_3 participants were significantly more locally biased in their initial five selections ($t(359), p < 0.001, d = 1.19$). Since they were already familiar with the *Location rule*, and it is probable they searched by ascending towards the maximum through small incremental steps. EE_3 participants had a significantly lower proportion of re-selections (0.19 in Experiment 3 vs 0.28 in Experiment 2) ($\chi^2(1, N = 1367) = 17.16, p < 0.001$). Given their higher performance scores, EE_3 participants were likely to have a strategy more adapted to the task than EE_2 participants, where participants were still learning the reward structure. Indeed, EE_2 participants had a tendency to settle on a sub-optimal tile, finding the maximum tile in 0.62 of grids. EE_3 participants took smaller exploratory steps but found the maximum in 0.81 of the grids ($\chi^2(1, N = 185) = 6.69, p = 0.01$).

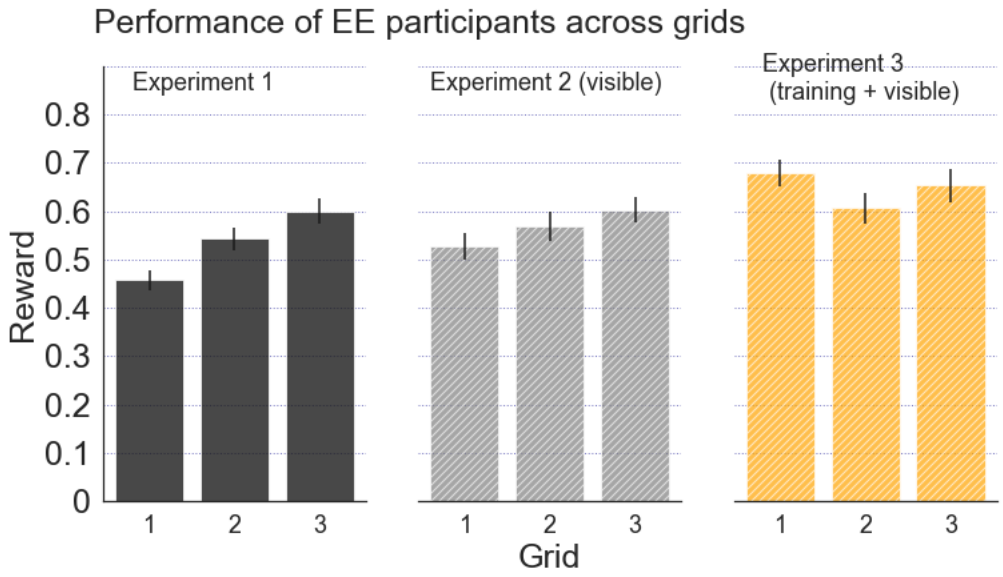


Figure 2.14: Average performance of EE participants across all three grids in Experiment 1, 2 and 3. The error bars show the standard error of the mean. The plots show significant progress in Experiment 1*, better initial performance* and slight progress in Experiment 2*, and a better overall performance but not progress across grids in Experiment 3.

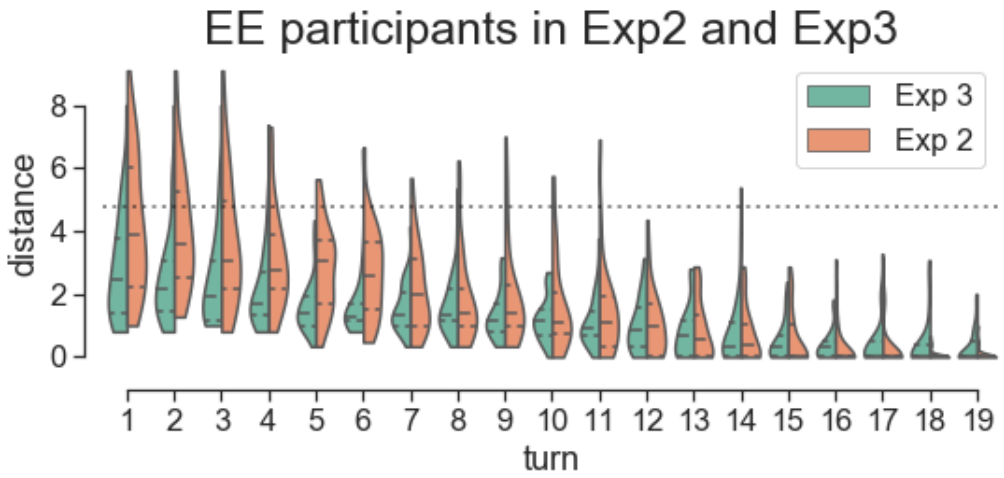


Figure 2.15: Distance between selections of participants (see Figure 2.8 for details). EE participants in Experiment 2 had more global observations than EE participants in Experiment 3. This can be explained by that fact that they had no prior knowledge about the task structure.

2.4.3 Interim discussion

We designed Experiment 3 to control for participants' strong exploratory drive. We hypothesised that participants' exploration was motivated by an intrinsic epistemic drive caused by uncertainty about the task structure. Contrary to our hypothesis, many participants still engaged in full exploratory behaviour despite being trained on the task structure prior to the task. Given this result, we hypothesised that participants might be motivated by observing new rewards rather than learning the underlying reward structure *per se* and that this effect might be emphasised when rewards remain visible after having been observed. Indeed, in Experiment 2, where rewards remained visible, significantly more participants engaged in full-exploratory behaviour than in Experiment 1. We designed Experiment 4 to account for these two factors of epistemic motivation:

1) wanting to learn about the underlying reward structure and 2) wanting to attend novel information.

2.5 Experiment 4

Experiment 4 followed the design details of Experiment 3, except that rewards were not displayed continuously after they had been selected - they are displayed on the tile and disappear after 1.5s, like in Experiment 1.

Our main hypothesis for Experiment 4 was that fewer participants would engage in *full exploratory* behaviour, since the epistemic reward is attenuated by not having the tiles visible after they have been selected and having training grids prior to the task. We predicted EE participants would perform similarly or slightly worse than in Experiment 3, because of the constraints of not having previous observations visible, but better than in Experiment 1 and 2. We also predicted we would observe little or no progress across grids.

2.5.1 Methods

39 participants were recruited using Amazon Mechanical Turk. In Experiment 1 and 2, the sample size was larger as participants were subsequently divided into two different conditions after the initial three grids (see Chapter 6). Participants were given the same instructions as in Experiment 3. Similarly, participants were paid a base reward of \$0.2 and a bonus of up to \$0.6 proportional to their performance. One participant was excluded for failing to select more than two different tiles, and one was excluded because their performance was worse than chance.

2.5.2 Results

In agreement with our hypothesis, only one participant out of 37 engaged in *Full Exploration*. This was (strikingly) less than in any other experiment. It supports the idea that participants' strategies were driven by an epistemic drive which was twofold:

First, participants were motivated to reveal the underlying reward structure, i.e., reducing the uncertainty about the structure of the task, or about the location of the maximum. Indeed, participants were less likely to engage in FE behaviour in Experiment 4 (known structure and disappearing observations) than Experiment 1 (unknown structure and disappearing observations), and significantly less in Experiment 3 (known structure and visible observations) than Experiment 2 (unknown structure and visible observations).

Second, participants were motivated to observe the outcomes of individual actions. In Experiment 1,2 and 3 a significant proportion of FE participants selected the maximum but consistently opted for selecting novel options rather than re-selecting a previous maximum observation, with a preference for actions that were local to their last one. Participants' drive to select novel actions was enhanced by the fact that information did not need to be kept in working memory. They were less engaged in FE behaviour in Experiment 1 (non-visible observations) than Experiment 2 (visible observations), and, similarly, less in Experiment 4 (non-visible observations) than Experiment 3 (visible observations). Though EE participants in Experiment 3 performed slightly better than in Experiment 4, this was not significant ($t(61) = 0.93, p = 0.35, d = 0.04$). Participants in Experiment 4 improved their average performance slightly across tasks ($b = 0.02, se = 0.007, p = 0.02$).

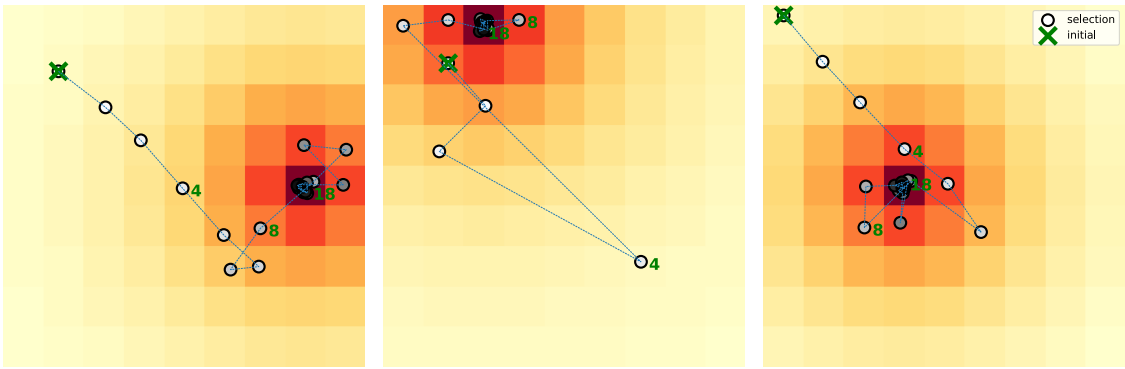


Figure 2.16: Selections of a participant from Experiment 4 across all three grids.

The average distance between the initial five exploratory selections of EE participants was not significantly different in Experiment 3 and Experiment 4 ($t(309) = -0.90, p = 0.37, d = -0.15$). EE participants in Experiment 4 explored significantly more locally than EE participants in Experiment 1 ($t(374) = -2.73, p = 0.007, d = 0.47$). Like in Experiment 3, this supports the hypothesis that participants who were familiar with the underlying structure of the grid were able to find the maximum by taking local exploratory steps until they eventually found the maximum. The selections of a participant in Experiment 4 are shown as an example of this in Figure 2.16.

2.6 Conclusion

In this chapter, we have focused on the behavioural analysis of participants across four experiments to study how people learn to select rewarding actions in a sequence of novel tasks. We found that some participants were able to learn the underlying structure while balancing exploration and exploitation to maximise their rewards across tasks. They improved their performance from one task to the next by transferring abstract knowledge about their environment.

However, we observed that a significant proportion of participants consistently engaged in purely exploratory behaviour across tasks, largely ignoring the reward incentive. We showed that this behaviour could be manipulated by controlling the availability of information as the learner selected actions, and by giving participants knowledge before they engaged with the task. We suggest that people are motivated by two types of epistemic drives: 1) to reduce uncertainty and learn about the structure of the task and 2) to observe new evidence, regardless of its informativeness about the global task structure. The latter was evident when participants continued valuing new actions over maximising rewards, even when they were familiar with the task structure.

Different potential mechanisms underpinning curiosity have been discussed in the literature, and could be connected to how people learn in new environments when combined with trying to achieve goals or maximising utility. One such strategy is to entirely dismiss reward feedback, giving rise to a strong novelty drive. This *novelty search* mechanism has been been shown to be very successful in the context of Evolutionary Strategies when learning policies for tasks with tricky reward functions (Lehman & Stanley, 2011). Dismissing rewards could be one of the exploratory strategies in the human adaptive toolbox. Additionally, some studies have shown that people are biased towards surprise (Gottlieb *et al.*, 2013; Itti & Baldi, 2006). Selecting new actions would make sense under the assumption of possible change, or if one believes that the environment is adversarial (i.e. trying to fool the learner). Third, the idea of *epistemic actions* could explain part of people's strong drive to select new actions, especially under the constraint of cognitive load, when keeping previous observations in memory is expensive or unrealistic. Epistemic actions refer to actions *in the world* that help solve problems by changing the mental state of the agent, as opposed to performing computations in the head (Kirsh & Maglio, 1994). An example of this behaviour is the use of sticky-notes, or of arranging documents in a way that makes it easier

to retrieve them rather than by memory alone. In the case of our experiment, observing new information might have been perceived as much cheaper than the possibility of generalising from few observations.

Recent accounts have presented the combination of uncertainty directed search and random exploration as the main mechanisms behind people's strategies (see e.g. Gershman, 2018; Schulz & Gershman, 2019). These results initially stem from research with experimental designs restrained to limited action-spaces (e.g. two- or four-armed bandits (Wilson *et al.*, 2014; Speekenbrink & Konstantinidis, 2015). More recent work has studied the behaviour of people on larger decision spaces (Schulz *et al.*, 2017b; Wu *et al.*, 2018) and concluded that the best account for human behaviour comes from a combination of a Gaussian Process model, to model human generalisation, and uncertainty directed search strategies. Our experiments show that the behaviour of participants largely deviates from the qualitative predictions of such models. Instead, novelty and local search seem to be prominent factors behind people's strategies. It seemed evident that many participants relied on the use of simple heuristic strategies rather than constructing a model of their environment to select actions that maximise information or rewards. We highlight that studying individual differences amongst participants can help us better understand the complex mechanisms at play during active learning in new environments. We hope that by pointing out surprising facets of human behaviour, this empirical study can guide the design of better computational models of human learning and exploration. In the next chapter, we develop a computational framework to gain further insight into people's representations and strategies when learning in new environments.

Chapter 3

Learning the structure of the world and simpler heuristics: Toward a general model of human exploration in vast decision spaces

3.1 Introduction

In the previous chapter, we found that people used diverse and qualitatively different strategies to maximise rewards in new environments. For example, participants differed in the type of search they conducted (e.g. *global* vs *local* search), in how much they explored, and in their overall performance. Across our four experiments, the type of strategy used by participants was influenced by the environment they were presented with. This was made evident by the group

level differences we observed. However, differences were also observed amongst participants within the same experimental context. In the first three experiments, we found systematic differences between subgroups of participants, both in how much they explored, and in their performance.

In Chapter 2, we categorised the types of behaviours solely according to people's propensity to explore. Though this was sufficient to highlight the important differences in behaviour observed across participants, the approach remains rather limited. Only looking at one metric (i.e. the explore-exploit ratio) does not lead to an understanding of the fine-grained differences and similarities between participants, nor the processes that can explain them. This chapter describes a modelling framework where the strategies used by participants can be explained in terms of their underlying processes.

Characterising individual differences is typically not a central aspect of the design of cognitive models. Often, cognitive modelling relies on data that is averaged or aggregated across subjects, and ignores individual variation (Navarro *et al.*, 2006). If the performance of participants truly is the same except for the noise, this has the potential benefit of removing the effects of the noise, thus yielding a more accurate representation of the underlying psychological phenomenon. If there are genuine differences amongst participants however, averaging the data can lead to misleading results and yield models that look nothing like individuals. Navarro *et al.* (2006) give as an example for this, an experiment where participants are asked which number is most unlucky. Depending on their culture, people might report numbers such as '13' (originally from a European tradition), '4' (4 is considered unlucky in Chinese tradition) or '87' (87 is considered an unlucky number for Australian cricket players). When averaging the data from the unlucky number experiment, we would get a number that was not given by any of the participants. Though the primary goal of cognitive modelling is to abstract

common cognitive processes, this simple example shows how important it is to also ask how people are different.

In the literature on explore-exploit type problems, much of the modelling effort has been concerned with comparing different classes of models to determine which model makes the best predictions about the data. A wide range of models have been examined in relationship to people's decision-making in such problems, ranging from diverse heuristics to more developed learning models. Typically, a model $\mathcal{M}$ consists in a set of structural assumptions $\mathcal{S}$, and in parameters θ whose meaning is specific to the model, given the choice of $\mathcal{S}$. In cognitive modelling, the structural assumptions $\mathcal{S}$ usually correspond to hypotheses about underlying cognitive processes (e.g., a given learning model, or a decision strategy). The task of studying individual differences can be difficult as differences can be expressed both in their variation in parametrisation of a model (θ) when fit individually to participants, but also in the class of model that best describes a given participant (see e.g., Borji & Itti, 2013; Steyvers *et al.*, 2009; Speekenbrink & Konstantinidis, 2015; Schulz *et al.*, 2017b).

In order to simplify this process, we develop a modelling framework within which the behaviour of participants can be represented and measured in a continuous parameter space. Rather than modelling participants using different classes of models, we design our modelling framework following the method of *model expansion*. Model expansion consists in designing a *general* model that includes the subset of *special case* models (e.g., a given heuristic, or a learning model). One motivation behind model expansion comes from the observation that scientific progress is driven by model checking and model revision, as opposed to through model comparison and model selection (Gelman & Shalizi, 2013). Many studies focus on the practice of model selection, which often assume the true model to be in the set of hypotheses considered. Unfortunately, this leads to very few studies

engaging in model checking, or reporting their process of model revision in their studies (for a discussion, see Gelman & Shalizi, 2013). We show here how the process of model expansion can help design richer and better predictive models to explain human behaviour, and is better inclined to the practice of model checking and model revision.

We first present desiderata for a general model to motivate our approach. Second, we describe our computational framework, and the psychological theories it carries. In a third part, we conduct an initial model check by focusing on both qualitative and quantitative aspects of model simulations, as well as the participant fits. We then conduct model checking against the experimental data and discuss some of the limits of the model. We suggest how it could be further expanded. Finally, we present a model-based analysis of the experimental data presented in the previous chapter data. Having introduced our modelling framework, we focus more specifically on the modelling of individual differences in the next chapter.

3.2 Desiderata for a general model of human search

We explain here what we believe to be the role of a cognitive model that attempts to account for human learning and decision making in reward maximising tasks, and how such a model can be evaluated.

We can separate models into two broad approaches: *scientific* and *technological* models. A scientific model aims to provide insight into understanding the "true" cognitive mechanisms that give rise to certain behaviours. Technological

models are more concerned with being able to either predict, control or imitate a given phenomena. While scientific models should also be evaluated along those measures, it reduces to a "technological" view if it limits itself to those goals. In contrast, a scientific model places more importance on its *explanatory value* and *interpretability* (Bernardo & Smith, 2009; Navarro, 2019). One problem with limiting the evaluation of a model to a close approximation of the empirical data is that it emphasises a focus on quantitative details while missing important qualitative patterns in the data. As Box (1976) famously remarked, "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad." **Qualitative model evaluations** are thus of primary importance, and particularly so in the psychological sciences where the relationship between the empirical data and the underlying cognitive processes remain poorly understood.

In designing a general model, we aim to have a model broad enough that it captures the different hypotheses at hand. Perhaps more ambitiously put, a general model should strive to **take into consideration the diverse theories available in the literature**. Unifying models have been responsible for important theoretical progress, resolving conflicts between contrasting theories. For example, Griffiths *et al.* (2007a) showed that exemplar and prototype models of categorisation (and everything in between) could be subsumed by a more general density estimation model based on the hierarchical Dirichlet process. Similarly, Lucas *et al.* (2015) showed that the conflicts between similarity-based models and rule-based accounts of human function learning could be captured by a unique model that combined their strengths.

In our case, we define our general model as a mixture of different components, with each carrying a specific hypothesis about a psychological process or strategy. When fit to a participant, the importance of each component is expressed through a given parameter of the model - and each participant behaviour is summarised

as a combination of these parameters. The model reduces the complexity of the data into **a succinct explanation of participant behaviour** given by the parametrisation of the different model components. This implies a trade-off between the flexibility of the model – allowing it to capture diverse hypotheses – and the simplicity of the model – constraining it to few, or ideally, a single explanation for a given behaviour.

If the parameters are cognitively meaningful, participant behaviour and model simulations should carry qualitative and quantitative similarities when parameter values coincide. The model should thus be checked for specific aspects, both qualitatively and quantitatively. For example, can we accurately explain the performance of participants, the distance between their selections, when different types of exploration occur across trials, or the degree to which they explore from the cognitive processes assumed in the model? Furthermore, if the model can capture differences that are empirically distinguishable, **the model should be recoverable**. This means that when fitting "fake" data, i.e. model simulations, the model should give back the generating parameters. To ensure that the number of free parameters used in the model is not larger than needed, a general model should yield **participant fits and predictions that are robustly better than simpler models** across participants. Indeed, successful model predictions on participant selections outside of the data the model was trained on guarantees the model is able to capture essential aspects of their behaviour.

3.2.1 Summary

In summary, we design our model with the aim of meeting the following criteria:

1. Flexible and capture a broad set of hypotheses
2. Offer succinct and interpretable explanations

3. Recoverable parameters

4. Identify distinct phenomena observed in the empirical data

5. Reproduce similar patterns of behaviour through simulations

6. Provide better participant fits and predictions than simpler models

In the next section, we outline and motivate the different components used in our model to capture different mechanisms of human behaviour in goal directed exploration, as studied in Chapter 2.

3.3 A general model of human decision making in explore-exploit problems

3.3.1 Model summary

Before presenting each model component in depth, we present in the box below a brief summary of the general model, its likelihood function and the different free parameters fitted to participants. The model consists of model based components, that rely on a Gaussian process model to capture the human ability to generalise from past observations and learn the structure of the world, and on simpler (model free) heuristic strategies. Model based strategies have the benefit of allowing for efficient and planned exploration, and often require less data, while model free strategies are less computationally expensive and can at times be more adaptive since they do not require an accurate representation of the environment to be effective.

General terms

$\mathbf{x}$: feature vector of an action e.g. location (x, y) and brightness of tile

y: reward/outcome of an action (point value)/

$a(\mathbf{x})$: model scores or acquisition function

$\{\mathbf{x}_n, y_n\}$: actions selected so far/observations

θ: Gaussian Process kernel hyper-parameters

General Model

$$a(\mathbf{x}) = \text{softmax}_\tau(\alpha \cdot \mu\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right) + \beta \cdot \sigma\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right) +$$
$$\nu \cdot \text{novelty}\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}\right) + \gamma \cdot \text{greedy}\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}\right) + \lambda \cdot \text{local-search}\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}\right))$$

Model weight parameters fit to individual participants

τ: softmax temperature (model confidence, or value sensitive actions)

α: GP expected mean, or reward driven actions.

β: GP variance, or uncertainty driven exploration.

γ: greedy re-selection of the maximum known value.

λ: local search component.

ν: random exploration, or novelty bonus.

Model likelihood

$$\mathcal{L}(\Theta|\mathbf{x}) = \text{softmax}_\tau(a(\mathbf{x}|\alpha, \beta, \gamma, \lambda, \nu))$$

with $\Theta = [\tau, \alpha, \beta, \gamma, \lambda, \nu]$.

3.3.2 Representing the world

In this section, we discuss the general computational problem of learning about the structure of the world while having to achieve goals within it. Having done this, we look at this problem more specifically by dividing it into two distinct problems: First, how do I represent the world (or task) and update my beliefs in light of evidence? Second, given my current beliefs about the world, what is my policy to achieve my goals?

One critical aspect of our experimental tasks was the large number of possible actions. Participants were presented with 81 possible choices and only had 20 selections in each grid. Much like in the real world, the outcome of an action (i.e. a tile) in the grids was not independent of the other available actions. Participants could learn the underlying structure of the task by predicting the outcome of unseen actions given their previous observations. In a first part, we describe our model of (*within-task*) generalisation in people, and how it relates to different search strategies.

When trying to maximise rewards, a participant might ask themselves: "What is the value of an action I haven't tried before?" This can be framed in an inductive way: "How can I generalise from my previous observations to unseen ones?". In

this case, the search for the most rewarding option is guided by knowledge about the abstract relationship between the different options (Gershman & Niv, 2015).

A standard Bayesian approach to predicting rewards would assume the learner begins with a *world model*, capturing *a priori* beliefs about the task structure (Courville *et al.*, 2006). Learning how the different features $\mathbf{x}$ of a given context relate to the rewards y is equivalent to the problem of learning a function $f(\mathbf{x})$. The model is specified only up to some set of unknown parameters, θ that specify some properties about $f(\mathbf{x})$, like which features are relevant to the task, or how they influence the expected reward of each action. Learning is interpreted as an attempt to recover the parameters of the generative model that wants to explain the observed events (i.e. the observed cues and outcomes, $\{\mathbf{x}_n, y_n\}_{n=1}^{N}$). We can encode background knowledge through a constrained space of hypotheses Θ, where each hypothesis θ represents a possible world structure that could explain the observed data. Finer-grained knowledge comes in the *prior probability* $P(\theta)$, the learner's degree of belief in a specific hypothesis prior to the observations. Bayes' rule (Bayes, 1763) updates priors to *posterior probabilities* $P(\theta|\{\mathbf{x}_n, y_n\})$ given the observed data $\{\mathbf{x}_n, y_n\}$:

$$P(\theta|\{\mathbf{x}_n, y_n\}) \propto P(\{\mathbf{x}_n, y_n\}|\theta)P(\theta)$$

.

We can exploit this probabilistic model for $f(\mathbf{x})$ when making decisions about the next observation, while integrating out uncertainty. Given such a model, the agent must decide how to act when maximizing long term reward. This is similar to the computational framework of Bayesian Optimisation (BO) (Snoek *et al.*, 2012), which consists of a model to represent the world and a (myopic) policy to act upon that representation. By constructing a model that takes into account all

previous observations of $f(\mathbf{x})$, we can find the maximum of difficult functions with relatively few evaluations when compared with methods that make use of local gradient or Hessian approximations. In the next two parts, we will first present how to construct such a probabilistic model and then the different policies that can be used to trade off between exploration and exploitation. We will discuss how these models can be useful to understand how people might learn, generalise and maximise reward across tasks.

3.3.3 Representing the world with Gaussian Processes

An agent with prior beliefs that match the underlying structure of the task will be much faster at finding the best action to take. The Gaussian process (GP) is a powerful and attractive prior distribution to express our assumptions about what kind of functions are plausible, since it does not require any arbitrary postulations about the parametric form of the unknown function (Williams & Rasmussen, 2006). The GP is defined by a collection of random variables, any finite number of which have a joint Gaussian distribution. The support and properties of this distribution are determined by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$, where

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$$
$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}')]$$

A key feature of GPs is that they enable us to be explicit about the different assumptions of our models, and their prior knowledge about the task. In contrast to neural network function approximators, they allow for psychologically interpretable parameters. For an overview of Gaussian processes, see (Williams & Rasmussen, 2006). A Gaussian Process model relies on the covariance function to

express a rich distribution over functions. In many applied problems, it is common practice to choose a general covariance function, like the Squared-Exponential kernel (SE) or Matérn $\frac{5}{2}$ kernel, to avoid restraining the space the of possible functions (Snoek *et al.*, 2012). The SE kernel is the most popular and default kernel for GPs and SVMs. This is due to its properties: it is universal, can be integrated against most functions and has only two parameters. Every function in its prior has infinitely many derivatives. However, sample functions with the SE co-variance function are sometimes seen as unrealistically smooth for practical optimization problems, so the Matérn $\frac{5}{2}$ kernel is sometimes preferred.

Figure 3.1 shows random samples from varying kernel functions (SE, exponential and linear kernels) to illustrate the different functional assumptions they can express. To draw a random function from a GP, we simply draw from the corresponding multivariate normal (with the co-variance matrix defined by the kernel function).

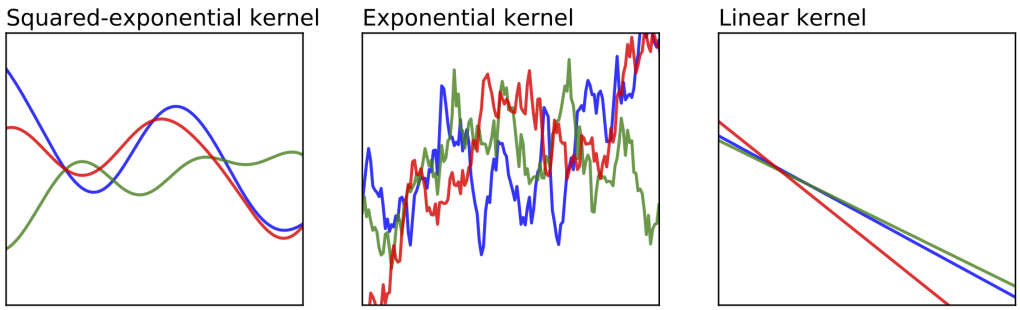


Figure 3.1: Samples from different kernel functions. These samples illustrate the different assumptions a kernel function can capture.

GPs have been have been successful at capturing human biases when making extrapolation judgements (Lucas *et al.*, 2015; Schulz *et al.*, 2016), unifying conflicting theories about how humans learn functions. Since then, they have been

used to model the human ability for generalisation in search tasks (Schulz *et al.*, 2017b; Wu *et al.*, 2017; Borji & Itti, 2013; Schulz *et al.*, 2018b). In Figure 3.2, we show how a GP might represent one of the grids presented to participants in the experimental framework, given the previous observations.

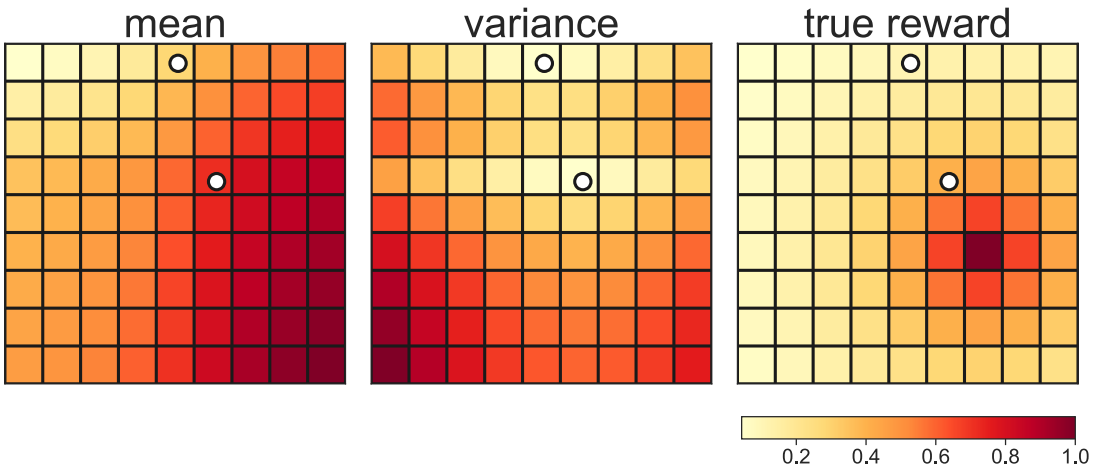


Figure 3.2: Expected mean and variance of GP model (integrated over different MCMC samples) conditional on previous observations. The rightmost grid shows the true reward structure. The white circles show two initial observations of a participant.

We use the Squared Exponential kernel (SE) in our model. The SE kernel assumes the same smoothness applies globally over the function. The SE kernel is governed by two parameters: the length-scale l and the variance σ^2. The length-scale accounts for the "wiggles" in the function, and how far it can extrapolate from the data. The variance defines the average distance between the function and its mean (basically a scale factor).

$$k_{SE} = \sigma^2 \; exp(-\frac{(x - x')^2}{2l^2})$$

To give an intuition for the effect of the length-scale on the kinds of hypotheses

considered likely by the GP, we show examples of different samples for different parameter values in the 1D and 2D case in Figure 3.3 and 3.4.

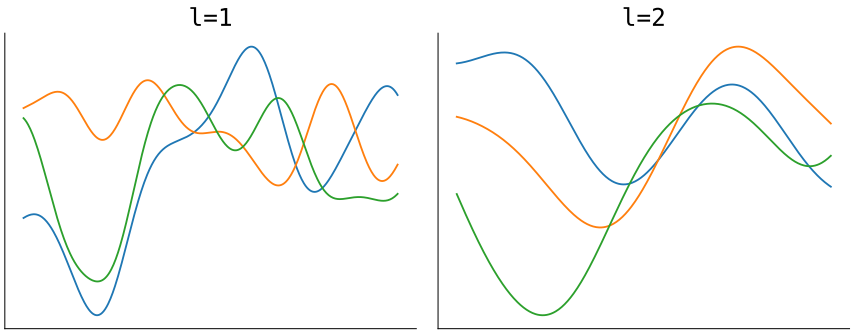


Figure 3.3: 1D samples from the SE kernel for different length-scale values

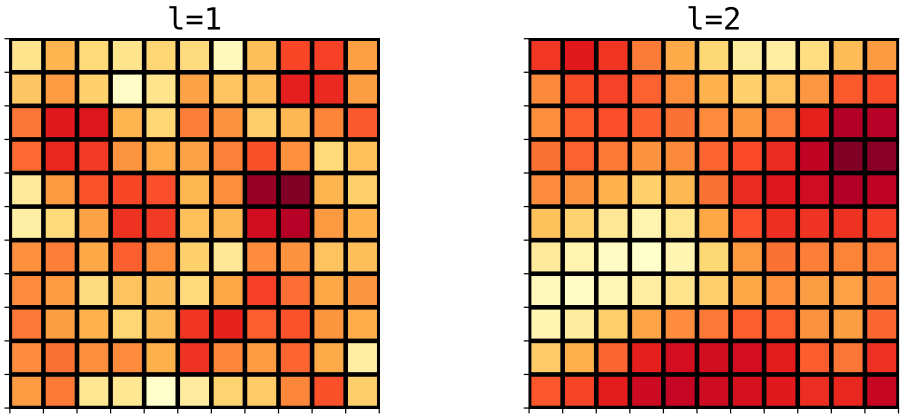


Figure 3.4: 2D samples from the SE kernel for different length-scale values.

We take a fully Bayesian treatment, as presented in Snoek *et al.* (2012), by getting samples from the posterior over GP hyper-parameters at each time step. These samples can be acquired efficiently using slice sampling, as shown by Murray & Adams (2010). This allow us to integrate out the hyper-parameter uncertainty out when computing the acquisition function score.

Each grid is considered independent of the other in our task, and the samples are

only taken conditional on the observations in the current grid. We discretise the GP predictions on the grids by taking the value at the centre of each tile.

3.3.4 Modelling guided search

The GP prior and the data observed so far induce a posterior over functions. In BO, the next action is selected by using this surrogate model of the unknown function (the posterior) to evaluate the value of each action. This surrogate model is represented in essence by the mean and variance of the GP. Minimising the regret over the complete optimal sequence of actions to find the maximum is typically computationally intractable. This has led to the introduction of many myopic heuristics (what we call acquisition functions), which are generally cheap to evaluate (Shahriari *et al.*, 2016). The acquisition function, which we denote by $a(\mathbf{x})$ decides what point in $\mathbf{x}$ should be evaluated next via a proxy optimization $\mathbf{x}_{\text{next}} = \text{argmax}_{\mathbf{x}}\, a(\mathbf{x})$. The acquisition function is a rule that determines how to resolve the exploration-exploitation trade-off. One strategy could be to select the tile with the highest variance, and hence to follow an explore-only strategy, whereby one would systematically pick the most uncertain option and acquire more knowledge about the task structure. On the other hand, a model selecting the highest expected reward would choose the tile that it believes will give the best reward. By failing to explore uncertain options, this strategy could easily make one fail to realise that a nearby choice offers a much higher reward. Selecting the appropriate strategy is a difficult problem given the lack of information about the function.

In general, these acquisition functions depend on the previous observations,as well as the GP hyper-parameters, $a\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right)$. There are a number of different strategies that aim to solve the explore-exploit problem by relying on the expected reward and variance under the GP model. For example, *Thompson*

sampling is a Bayesian algorithm for sequential decision-making that consists of probability matching on the posterior probability that an option is best. In other words, the selection of an action is proportional to the probability of it being the most rewarding action. Thompson sampling has strong optimality guarantees in sequential decision-making problems and has been shown to have state-of-the-art empirical performance in many domains (Chapelle & Li, 2011; Kaufmann *et al.*, 2012). Another popular strategy is the Upper Confidence Bound (UCB) (Srinivas *et al.*, 2009). UCB is an explicit trade-off between exploration (the variance function $\sigma^2(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)$) and exploitation function (the expected reward $\mu(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)$):

$$a_{\mathrm{UCB}}\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right) = \mu\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right) + \beta\sigma\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right)$$

It selects the option with the upper confidence bound, with β being defined depending on context. Intuitively, UCB corresponds to select actions optimistically following the assumption that the most uncertain actions have the potential to be the most rewarding. The algorithm has been shown to perform efficiently in a limited number of samples to find the global optimum of many multimodal black-box functions (Srinivas *et al.*, 2009).

Another popular and successful strategy is to maximize the Expected Improvement(EI) over the current best (Jones *et al.*, 1998). This can be computed analytically under the Gaussian process as:

$$a_{\mathrm{E1}}\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right) = \sigma\left(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta\right)\left(\gamma(\mathbf{x})\Phi(\gamma(\mathbf{x})) + \mathcal{N}(\gamma(\mathbf{x}); 0, 1)\right)$$

Figure 3.5 shows the different steps of Bayesian Optimisation, and the contrasting predictions made by the different acquisition functions discussed above.

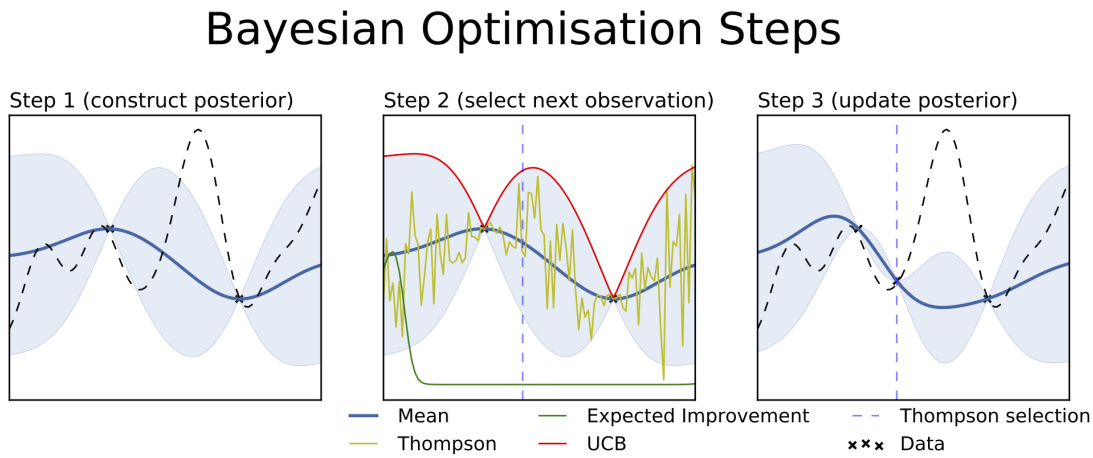


Figure 3.5: The different steps that define Bayesian Optimisation. Step 2 shows the contrasting evaluations of three different acquisition functions: UCB, Thompson sampling and Expected Improvement.

These are just three of the many different acquisition functions in the literature, with more complex acquisition functions existing, such as entropy based ones, or strategies involving a portfolio of acquisition functions (for a review, see Shahriari *et al.*, 2016). While one could be tempted to compare their distinct predictions on participant decisions, many acquisition functions lack clear psychological interpretations, and our main goal is to ask: Do people use generalisation to guide their search? Do people rely on a measure of uncertainty when they explore? To answer these two questions, our model includes of a weighting parameter α to capture *utility driven actions* (the expected reward under the GP $\mu(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)$), and a β parameter as an explicit *uncertainty-driven exploration* parameter (the variance under the GP$\sigma^2(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)$) without further consideration of alternative transformations of the GP predictions.

3.3.5 Heuristics and biases in human search

Probabilistic models have been able to explain many aspects of human cognitive phenomena, like how people are able to successfully combine their estimations of perceptual uncertainty with prior knowledge (Kording & Wolpert, 2004; Tassinari *et al.*, 2006) or how people learn how to represent uncertain environments in specific cognitive tasks (Griffiths *et al.*, 2007b; Kemp & Tenenbaum, 2008). Despite these successes these models have been met with scepticism, primarily originating from two lines of thought. First, copious evidence has pointed out an evident gap between human behaviour and probabilistic reasoning (Kahneman & Tversky, 1982). Second, the computational complexity of probabilistic models largely surpass the brain's ability to compute for optimal solutions, motivating a need for short-cuts and heuristics (Anderson, 1991; Simon, 1955).

A large branch of research in the cognitive sciences has been concerned with the heuristics and biases employed by people as tractable solutions to the problems faced in the real world given the limited computational resources of the mind/brain. This is achieved by reducing the amount of information in the input data (i.e. only attending relevant features of the problem), or constraining the hypothesis space. The computational problems people are often faced with are so-called *inductive problems*, where people have to infer a plausible structure from limited data. A number of studies have thus looked at the kinds of representations and algorithms people might use that allow for efficient, yet computationally tractable inference (Sanborn *et al.*, 2010; Bramley *et al.*, 2017; Griffiths *et al.*, 2015; Daw & Courville, 2008; Bonawitz *et al.*, 2014; Lieder *et al.*, 2018, 2014).

Beyond simplifying the problem, research studies at the algorithmic level of analysis have considered low-resource policies that aren't necessarily about representation *per se*. In this context, heuristics have been reported to, at times, outperform

full-information strategies and more complex models (Gigerenzer, 2008; Parpart *et al.*, 2018).

In light of this, and inspired by models existing in the literature as well as patterns observed in participants in our experiments, we consider three low-resource policies as components to our general model.

Undirected exploration

So far, we have introduced two forms of directed search: Uncertainty driven search, which adds a bonus to actions based on an agent's uncertainty about unseen values, and utility driven search, which selects actions according to their expected reward. We presented acquisitition functions that combine these two types of directed exploration, e.g. EI, UCB or Thompson sampling. Beyond directed exploration, there is also significant evidence in the literature for random exploration in human search. In fact, recent theories present human exploration as a combination of directed search and random exploration (see, e.g. Wilson *et al.*, 2014; Gershman, 2018; Schulz & Gershman, 2019). Random exploration is perhaps most simply explained with the ϵ-greedy algorithm, one of the first and simplest reinforcement learning algorithms. It ignores the value of information entirely and explores by selecting random actions with probability ϵ and maximising with probability 1-ϵ. By gradually decreasing ϵ (the agent's propensity to explore), the agent will eventually learn the values of the different actions and select the optimal action.

$$novelty(\mathbf{x}) = \begin{cases} 0, & \text{if } x_{observed} \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

In our model, we can express random exploration as a uniform probability of choosing novel tiles. In our mixture of policies, this can be formalised by assigning a score of one to all previously un-visited tiles; because of the deterministic nature of our grids, re-selecting an observed tile should not be characterised as an exploratory action. This novelty component thus has a dual interpretation 1) an inclination to favour new actions and 2) an undirected search strategy, i.e. a non-preferential way of exploring new actions (Schulz *et al.*, 2017a).

ϵ-greedy strategy

We introduced the idea of low-cost policies. The simplest, and perhaps most famous policy in RL problems is the ϵ-greedy algorithm. It consists in a trade-off between random exploration (discussed in the previous section) and the greedy re-selection of the maximum known value. Despite completely ignoring the structure of the problem, it has been shown to do well in many problems, and after enough exploration will eventually find the best solution. In support of this greedy strategy, we define a weight component that assigns a probability weight to re-select the (last observed) maximum known value. This simple *greedy* policy can be described as follows:

$$greedy(\mathbf{x}) = \begin{cases} 1, & \text{if } \max(\mathbf{x}_n) \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Greediness is a popular term when describing cheap solutions to complex problems. An example strategy that could be described as greedy is the "win-stay, lose-sample" heuristic, a decision making strategy which has been used to explain how people might sample for hypotheses in the domain of causal learning (Bonawitz *et al.*, 2014). While myopic and locally optimal ("greedy") strategies are very

relevant to our study, we restrict our use of the word greedy to the ϵ-greedy sense here.

Local search heuristic

In Chapter 2, we reported a consistent local bias in how participants selected new actions: they displayed a preference for nearby actions. We noted that this was also reported in the experimental data of Wu *et al.* (2017) in grid tasks similar to our experiments presented in Chapter 2, and may correspond to phenomena in other domains such as causal learning (Bramley *et al.*, 2015) and category learning (Markant *et al.*, 2016b). A wide range of robust optimisation methods, such as gradient based methods, rely on local search (Ruder, 2016) and can be used in combination with global optimisation methods for significant improvements in terms of efficiency of search (e.g. McLeod *et al.*, 2018; Acerbi & Ji, 2017).

To model a local search heuristic, we use the inverse Manhattan distance (IMD) to the last observation. We choose the IMD to remain agnostic to the reward function and too reflect the grid structure of our task.

$$\mathrm{IMD}\left(\mathbf{x}, \mathbf{x}'\right) = \frac{1}{\sum_{i=1}^{n} |x_i - x_i'|}$$

For the special case where $\mathbf{x} = \mathbf{x}'$, we set $\mathrm{IMD}\left(\mathbf{x}, \mathbf{x}'\right) = 1$. For the special case where $\mathbf{x} = \mathbf{x}'$, we set $\mathrm{IMD}\left(\mathbf{x}, \mathbf{x}'\right) = 1$.

To control for the smoothness of the localisation bias, we transform the distances through a softmax function. For simplicity, and to avoid model over-specification, we use a fixed decay parameter. We found that fitting the distance decay temperature to participants led to mismatches when doing model recovery, due to correlations with the importance of the weight component, and with the softmax

temperature[1]. We use a temperature parameter of 0.5 to account for a preference for neighbouring tiles without restraining it to the direct neighbours. A small temperature of e.g. 0.001 means no local bias (almost uniform weight on all tiles), while a larger value, e.g. 1, means most of the probability mass is put on the tiles directly adjacent to the previous selection (see Figure 3.6 for a visual intuition).

We set the score of returning to the previous observation to 0, given that we would consider a return to a previously observed tile as an exploitative choice.

$$\text{local-search}(\mathbf{x}) = \text{softmax}_{0.5}(\text{IMD}(\mathbf{x}, x_{n-1})) \tag{3.3}$$

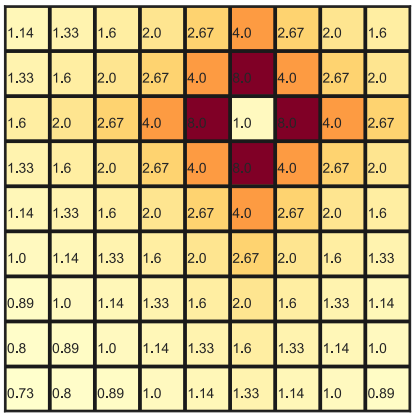
3.3.6 Value sensitive exploration and noisy behaviour

In the model, we transform the score of the mixture of components $a(\mathbf{x})$ into probabilities by using the softmax choice rule with inverse temperature parameter τ:

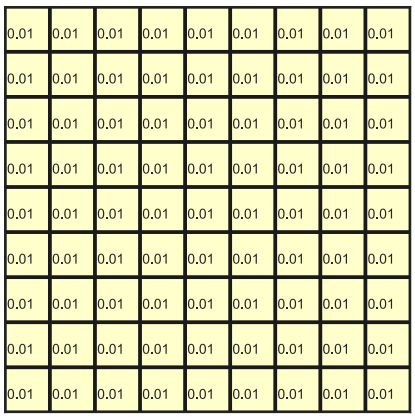
$$\text{softmax}_{\tau}(a(\mathbf{x}) = \frac{\exp(a(\mathbf{x})/\tau)}{\sum_j \exp(a(x_j)/\tau)} \tag{3.4}$$

The softmax rule is a popular decision rule in the decision-making literature, due to its simplicity, its biological plausibility and empirical support (Collins & Frank, 2014; Daw *et al.*, 2006; Schulz & Gershman, 2019; Speekenbrink & Konstantinidis, 2015). Daw *et al.* (2006) describe it as follows: "With softmax, the decision to explore and the choice of which suboptimal action to take are determined probabilistically on the basis of the actions' relative expected values.".

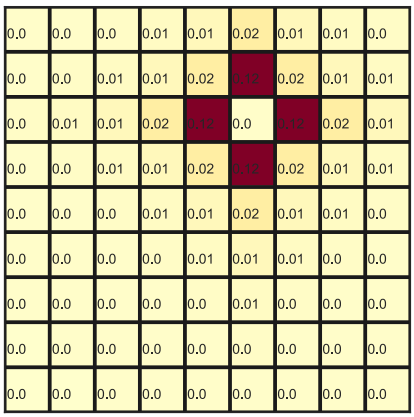
[1]To give an intuition for this, a small softmax temperature (e.g. 0.001) and a weak local bias (e.g. 0.1) would predict very similarly to a model with a more noisy softmax temperature (e.g. 0.05) and a much stronger local bias (e.g. 1).



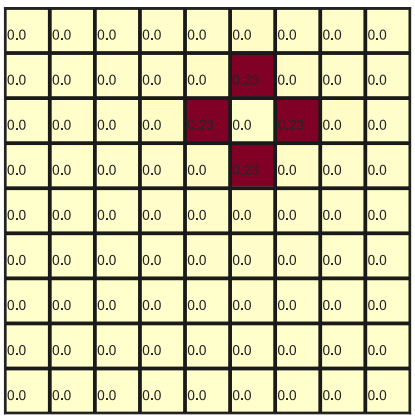
(a) Heatmap of Inverse Manhattan distances to the previous selection (x_{n-1}).



(b) Heatmap of IMD with softmax temperature $\tau = 0.001$



(c) softmax temperature $\tau = 0.5$



(d) softmax temperature $\tau = 1$

Figure 3.6: Grid visualisations showing the Inverse Manhattan Distances and the reward shape of the local search component for different softmax temperature parameters.

Daw *et al.* (2006) further characterise it as *value-sensitive* exploration. In their study, the softmax is put in contrast with *undirected* exploration, as implemented through the ϵ-greedy model. Following them, we adopt here the definition of *value-sensitive* exploration for exploratory behaviour as modulated by softmax temperature. Indeed, for small values of τ, small value differences lead to strong action probability differences. For large values of τ, the model largely ignores the predicted values by different model components and simply selects options at random, without differentiating exploration and exploitation. It is important to note that this is different from *undirected exploration*, like under the ϵ-greedy model, which picks any option that *is not* the max-known equiprobably (see Figure 3.7 for a visual explanation). In this case, there is an explicit trade-off between exploration and exploitation, but with no preferences for what to pick in terms of novel, or uncertain, options. In the future, it is in this sense that we refer to *undirected exploration*.

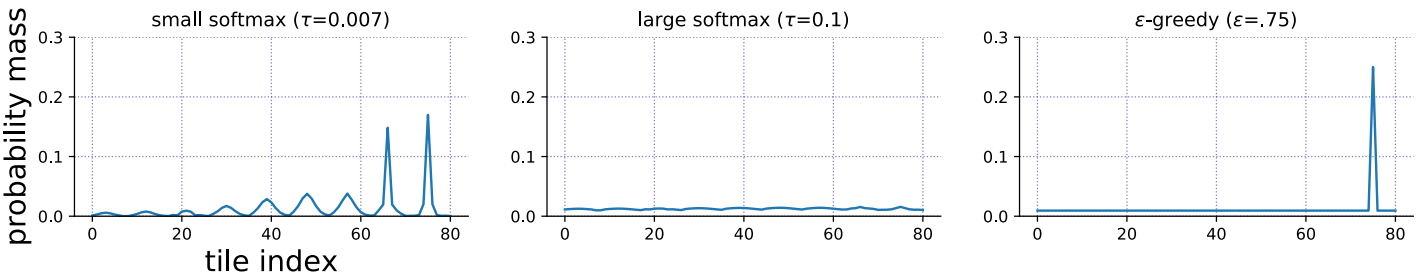


Figure 3.7: These plots show the probability scores attributed by the model for small and large softmax temperature parameters, and for an ϵ-greedy model. We highlight the difference in predctions made by the large softmax temperature and the ϵ-greedy model. ϵ-greedy implements *undirected exploration*, whereas larger values of τ lead to random uniform predictions. Smaller values of τ account for value-sensitive exploration.

Though high softmax temperatures have been associated with random exploration

(Schulz & Gershman, 2019; Wu *et al.*, 2018), it is more accurate to say that the softmax characterises random or unpredictable *behaviour*. A high softmax temperature parameter can generally be attributed to two cases: 1) the model captures the beliefs of a participant poorly, or 2) the participants is selecting at random. In general, the softmax temperature can be interpreted as the model's confidence in predicting an agent's behaviour given a set of beliefs about the underlying structure of the environment. In the next section, we explain how we fit our general model to participants.

3.3.7 Model fitting and model comparison

We fit models to individual participants by using a Differential Evolution algorithm (Storn & Price, 1997) to maximise the maximum likelihood function, followed by a gradient based optimisation step. We use the L1 penalty on all weight parameters. From a Bayesian perspective, the L1 penalty can be understood as a sparse prior for model components that favours setting non-significant components to zero. In this Bayesian framing, the maximum likelihood estimate (MLE) corresponds to the maximum a posteriori estimation (MAP).

With a general model, we hope to understand how different strategies interact for each participants, and better understand the salient differences in strategies across participants described in our experimental results. One of the aims of a general model is for it to offer interpretable explanations over different mechanisms for exploration. Here, each model component represents a distinct behaviour and mechanism during search. For better interpretability of the model, we normalise the weight parameters (i.e. all parameters but the softmax) to get their relative contributions. To ensure that the different parameters are representative of their contribution to the model predictions when weighted together, we rescale each component scores over their total range across all predictions.

To illustrate the need for rescaling over the total range, imagine a model with equal probability weight on the max known tile, and the four neighbours to the previous selection. The greedy component will put a score of 1 on the max known, while the (extreme) local bias term will put 0.25 on each neighbour. To put an equal weight on the five actions the model needs to give four times more importance to the local bias component than the greedy weight. Rescaling by the range of each component allows for more intuitive parameters, and would in the example above yield equal contributions from the greedy and local bias components.

We did not transform each component output into probabilities at every time step as this would not take the inter-dependence of scores across observations into account. For example, in the case of the variance predicted by the GP model, the scores might be very high early on in a grid, when only few observations have been made, but would be distinctly lower at the end when many tiles have been observed. Having the prediction scores transformed into probabilities at each time step would only keep the relative difference between the different tiles as information, but not the relative uncertainty over the different tiles as it progresses over the sequence of observations.

3.4 Model recovery of specific strategy types

As a first evaluation step for our model, we perform parameter recovery for simulated data generated by "special case" models. In our analysis, we focus on the parameters recovered by the general model and their interpretability i.e. how do the recovered parameters relate to, and explain, the behaviours exhibited by the "special case" model simulations. We focus on explaining three types of behaviours that are popular in the literature. 1) we simulate from an ϵ-greedy model ($\epsilon = 0.3$ and $\epsilon = 0.5$). We chose two distinct parametrisation to ensure the

model could reliably capture the correct ϵ value. We also study simulations from 2) a GP-UCB model ($\alpha = 1$ and $\beta = 0.8$) and 3) a heuristic we name "line-search", inspired by some of the participant behaviour observed in our experiments. The aim of this analysis is to evaluate the ability of our general model to identify and distinguish qualitatively different patterns of behaviour in an interpretable way (cf. desiderata 1, 2 and 4, see Section 3.2). We focus on the ability to reproduce qualitative patterns of behaviour (desideratum 5), the recoverability of parameters from simulations (3), the model's predictive ability and robustness when fit to actual participant data (6) in Section 3.5.

3.4.1 ϵ-greedy (ϵ=0.5)

For all special case models, we ran 71 simulations on the grids presented to participants from Experiment 1, corresponding to the number of participants in Experiment 1. We then proceeded to fit the general model to the special case simulations. Figure 3.8 shows the selections of one of the model simulations across the three grids, where the model balances random exploration and exploitation of the max known tile.

When looking at the recovered parameter weights, the greedy component was the highest contributor ($M = 0.71, SD = 0.15$) (see Figure 3.9). Because of the expressivity of the model, some of the other components inevitably capture some of the random exploration of the simulations. The largest non-greedy weight is the novelty component ($M = 0.10, SD = 0.13$), as it attributes a uniform bonus to non-explored tiles, much like the ϵ exploration term of the ϵ greedy algorithm. The average contributions of the other parameters were all below 0.1.

To better understand the fit parameters and how they relate to the model predictions, we look at the recovered parameters for a single simulation. Figure 3.10

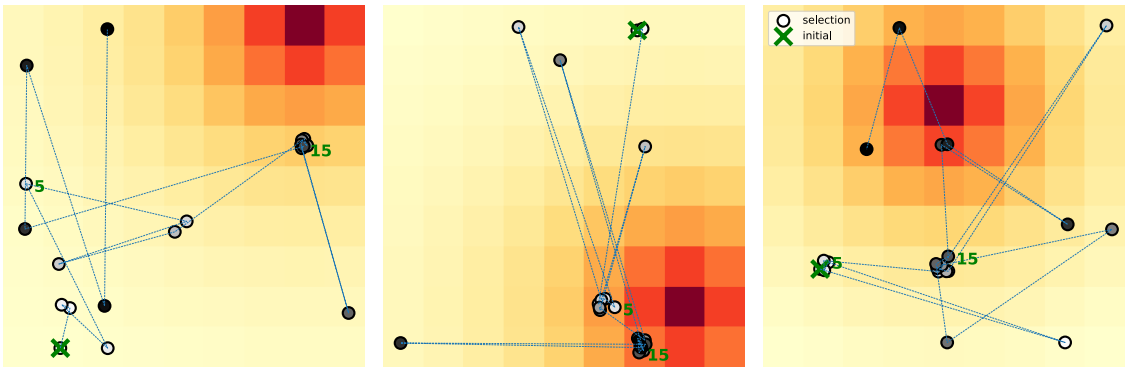


Figure 3.8: ϵ-greedy (ϵ=0.5) model simulation. The green cross marks the initial observation. Markers indicate observations, and a darker shade means an observation was later in the trial. Numbers indicate the index of the closest observation marker. The colour of a tile indicates its associated reward value (the darkest is the maximum).

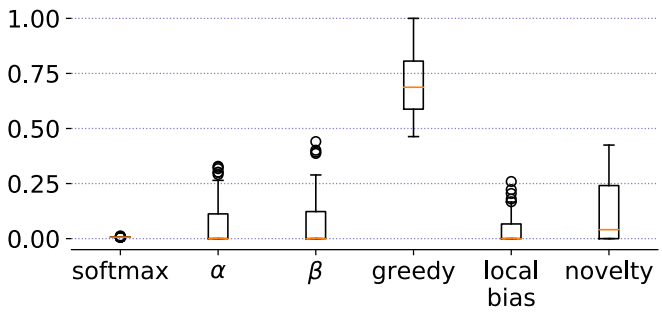


Figure 3.9: Recovered parameters for ϵ-greedy (ϵ=0.5) simulations under the general model.

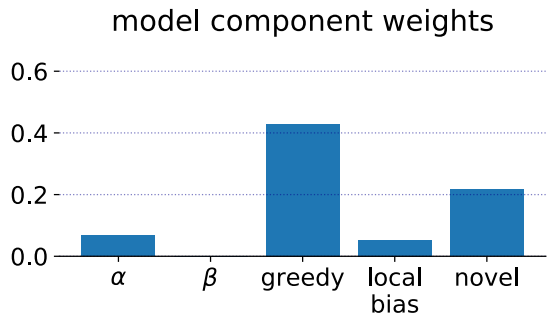


Figure 3.10: Parameter weights for ϵ-greedy (ϵ=0.5) simulation under the general model.

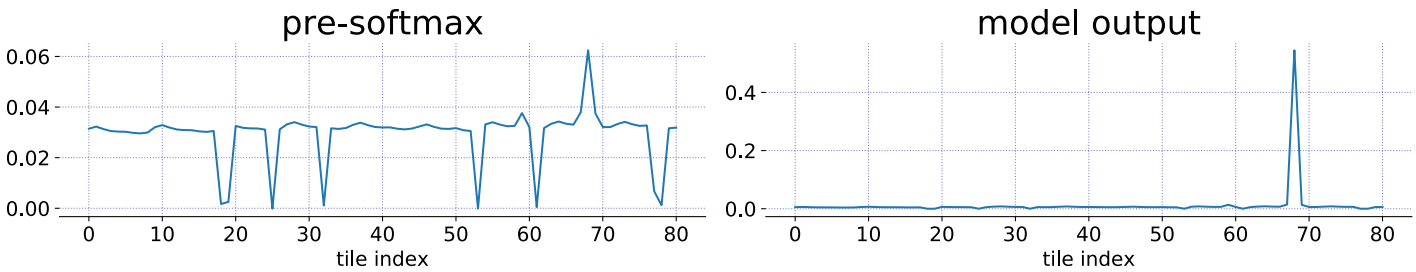


Figure 3.11: Combined scores of weighted model components and probability choices after the softmax choice rule is applied on trial 15 of an ϵ-greedy (ϵ=0.5) simulation.

shows the normalised contributions of the parameter weights. Of note, we see that local bias and the expected reward term have non-zero contributions. Figure 3.11 shows how the different components contribute to the predictions of the model, and the effect of the softmax transformation. The predictive scores of the *greedy*, *local bias* and *novelty* components are shown in Figure 3.12. The non-zero contributions are drastically reduced by the softmax ($\tau = 0.007$), which puts close to .5 of its probability mass on the max known (as would be expected for this parametrisation of the model). Across all simulations for ϵ=0.5, the average probability mass put on the max known was 0.50 (SD=0.06).

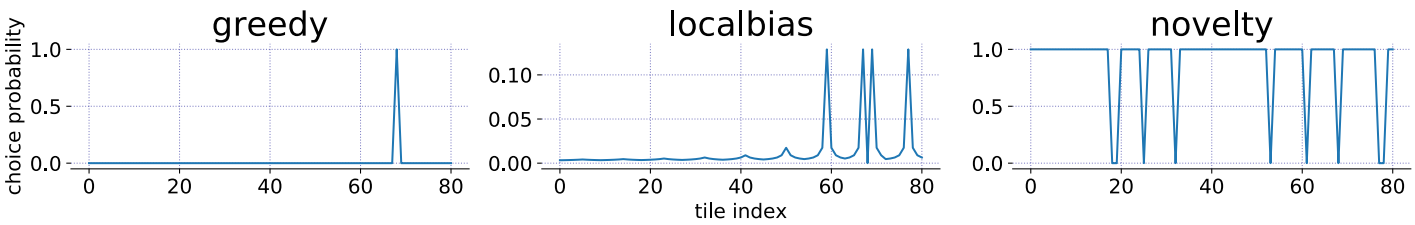


Figure 3.12: Prediction scores of individual components of general model on trial 15 of an ϵ-greedy (ϵ=0.5) simulation.

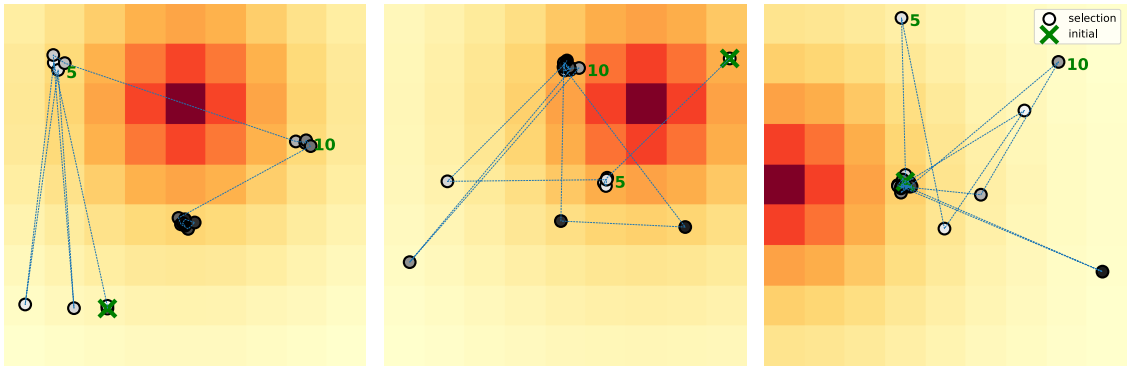


Figure 3.13: ϵ-greedy (ϵ=0.3) model simulation.

3.4.2 ϵ-greedy (ϵ=0.3)

As expected, similar results were obtained for ϵ-greedy simulations with ϵ=0.3. Figure 3.13 shows the selections of one of the model simulations across the three grids.

The parameter fits were again largely dominated by the *greedy* term, combined with small values for the softmax τ parameter (see Figure 3.14).

We inspect the parameters and predictions for a single simulation again. In this case again, the parameter contributions are dominated by the *greedy* term, with some weight contributions from the *novelty* and β components (see Figure 3.15).

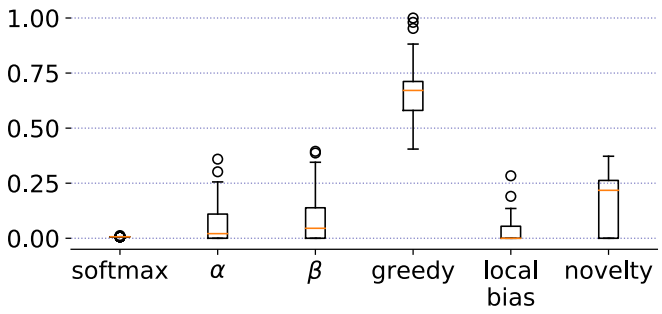


Figure 3.14: Recovered parameters for ϵ-greedy (ϵ=0.3) simulation under the general model.

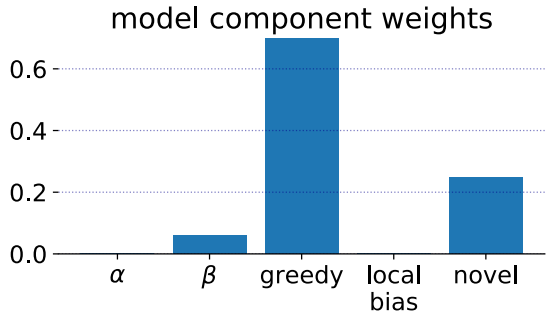


Figure 3.15: Parameter weights for ϵ-greedy (ϵ=0.3) simulation under the general model.

Like with the ϵ=0.5 example, these are largely squashed by the softmax. Most notably, this time the model puts around 0.75 of the probability mass on the max known, which corresponds approximately to the generating simulation parameter ϵ=0.3 (see Figure 3.16). When looking across all model simulations, the general model put 0.70 probability mass on the max known across all selections (i.e. 71 models and 60 turns) (SD=0.01).

In summary, we find that the recovered parameters were coherent with the behaviour of the ϵ-greedy algorithm, namely greedy re-selection of the max known, but no preference for the exploratory strategies in the model. We also find that

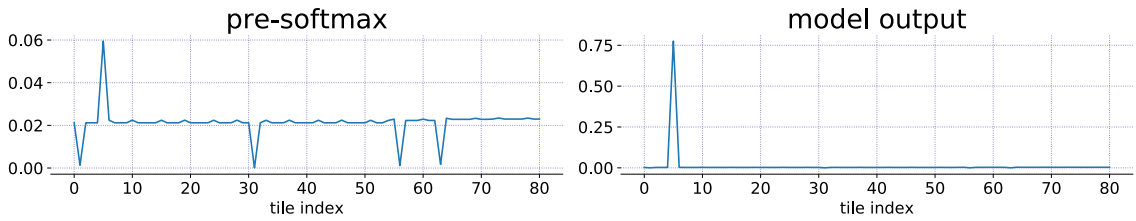


Figure 3.16: Combined scores of weighted model components and probability choices after the softmax choice rule is applied on trial 15 of an ϵ-greedy (ϵ=0.3) simulation.

the amount of exploration could be recovered by evaluating the predicted weight on the maximum known. In both cases, it matched the true parametrisation of the ϵ-greedy simulations.

3.4.3 Line-search heuristic

The next model simulations we study are inspired by some of the behaviour observed in our experiments, particularly by some of the *Full Explore* participants. In Chapter 2, we noted that some participants explored by choosing local gradient-ascent steps, essentially creating lines of observations to ascend to the maximum value. To mimic this behaviour, we implement a heuristic model that put probability mass on 1) the next point in the direction of the two previous observations (i.e. either horizontal or vertical, and in the same direction), 2) on the orthogonal tiles from the last observation, 3) on the neighbours of the maximum, and 4) a random exploration term. This heuristics model differs slightly from what people do, as they tended to specifically select lines that ascend reward gradients. We predict the general model will be able to capture this gradient ascent behaviour in humans through a combination of local search and the expected rewards of the Gaussian Process model.

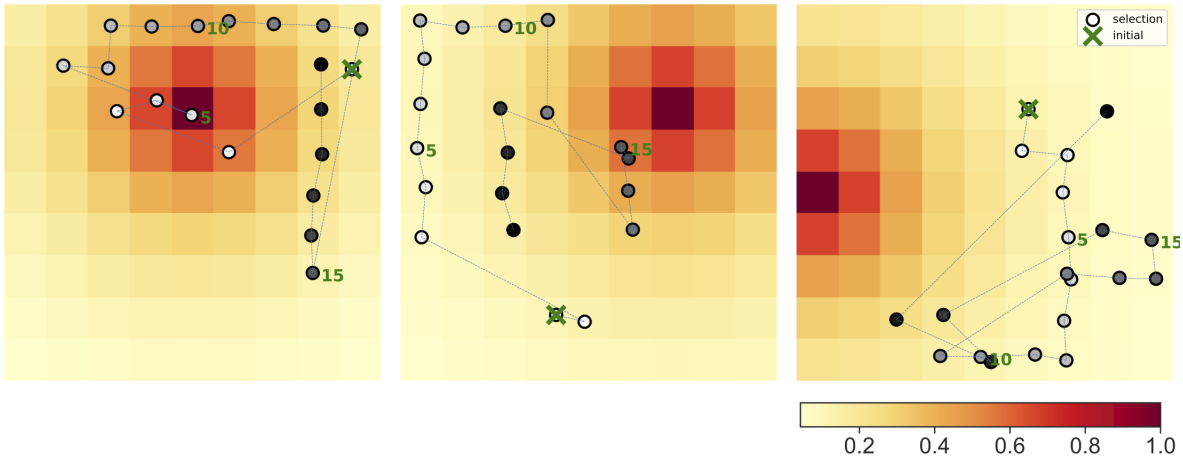


Figure 3.17: Line-search heuristic model simulation. The green cross marks the initial observation. Markers indicate observations, and a darker shade means an observation was later in the trial. Numbers indicate the index of the closest observation marker. The colour of a tile indicates its associated reward value (the darkest is the maximum).

We discuss this algorithm and how well it explains participant behaviour in more details in Section 3.6. A simulation from this model can be seen in Figure 3.17.

Again, we fit the general model to simulations from the line-search heuristic model. In this case, the simulations were mostly recovered by the *local bias* ($M = 0.50, SD = 0.10$) and *novelty* ($M = 0.39, SD = 0.12$) components of the general model, corresponding neatly to the observed behaviour of the model.

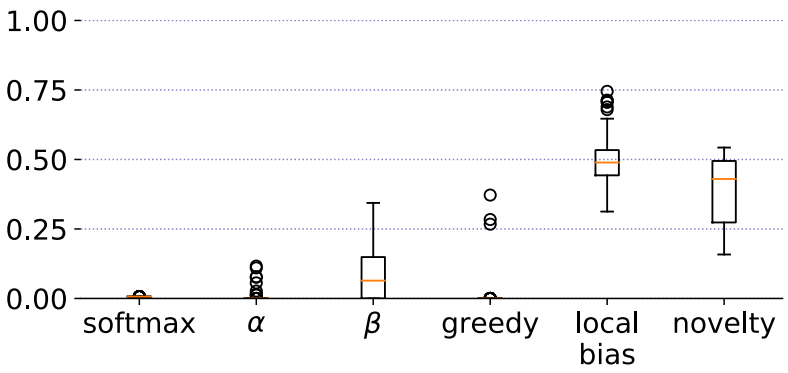


Figure 3.18: Recovered parameters for line-search simulation under the general model.

Indeed, the line-search agent explored (almost exclusively) while favouring neighbouring tiles. Again, we see that the model was able to recover parameters that were in line with the observed behaviour for another model that was not directly a sub-case of the general model.

3.4.4 GP-UCB

The final special case we study is a vanilla GP-UCB model simulation ($\alpha = 1, \beta = 0.8$) (see Figure 3.19). This model is a popular optimisation model in the Machine Learning literature Snoek *et al.* (2012); Shahriari *et al.* (2016) and has also been used to model human search in function optimisation and in grid-tasks Borji & Itti (2013); Schulz *et al.* (2017b); Wu *et al.* (2018). This model is directly a special case under the general model. Unsurprisingly, it was recovered neatly, with α ($M = 0.66, SD = 0.02$) and β ($M = 0.34, SD = 0.22$) parameters as the only contributing components (see Figure 3.20).

In conclusion, we have shown that our general model is able to recover and

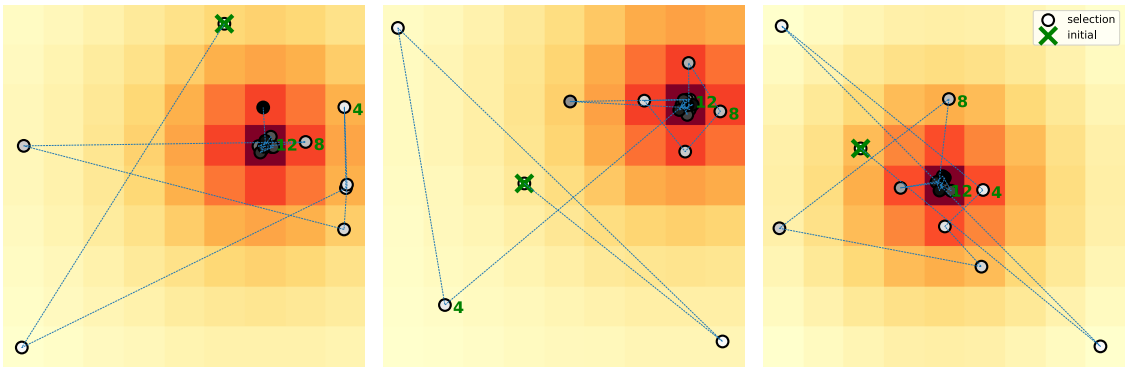


Figure 3.19: GP-UCB ($\beta = 0.8$) model simulation.

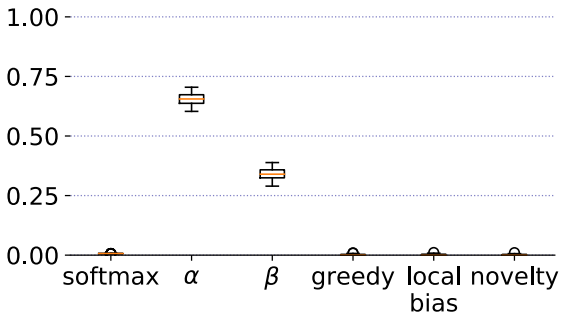


Figure 3.20: Recovered (normalised) parameters for GP-UCB model under the general model. The generating parameters were $\alpha = 1$ (the expected mean of the GP) and $\beta = 0.8$ (the variance predicted by the GP). The mismatch with the recovered parameters here is due to the normalisation over the complete range of $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ from the GP model.

provide interpretable parameters for three types of distinct search behaviours. Perhaps most importantly, this was the case when the model simulations were not generated by the general model itself, providing evidence for its broad coverage of behaviours. In the next part, we test this capacity further by fitting the model to participants. We then simulate the model forward based on the participant fits and attempt to recover them. This is to ensure that the model is well-specified, specifically that it is able to uniquely identify diverse types of behaviours.

3.5 Model fitting of participant data and model simulations

In this section we continue the evaluation of the model to assess its validity and robustness (desiderata 3 and 4). Our analyses focus in this part on participants in Experiment 1 (71 participants) as we considered it represented a large and diverse enough set of participant behaviours for the evaluation of the model.

3.5.1 Parameter recovery with model simulations

We perform parameter recovery to ensure each model parametrisation is meaningful, meaning that it captures distinct behavioural features. Parameter recovery consists in 1) fitting the general model to participants, 2) simulating data from the parameter estimates of participant behaviour (i.e. generating "fake" participant data) and 3) fitting the simulations with the same generating model (Wilson & Collins, 2019). A weak correlation between the simulated and recovered parameters would help point elucidate potential failures or biases in the model (see e.g., Nilsson *et al.*, 2011). This could stem from multiple causes, e.g. non linearly

separable parameters, or simply because the task does not produce diagnostic data.

We find that in all cases, the recovered parameters were highly correlated to the generating ones. These results support the existence of the different components implemented in the model as unique and independent behavioural characteristics of human exploratory search. These are: 1) "greedy" re-selection of the max known action, 2) utility driven actions (maximum expected value), 3) uncertainty driven exploration, 4) local search, and 5) (undirected) novelty driven search.

In Figure 3.21, we plot the correlation between the generating parameters and the recovered parameters with the generating parameters on the x-axis, and the recovered parameters on the y-axis. The generating parameters used were the MLE estimates on all three grids of participants in Experiment 1. We report rank correlation using Kendall's tau (r_τ), not to be confused with the softmax temperature parameter τ.

The rank-correlation between the generating and the recovered expected utility α parameters was $r_\tau = 0.70, p < 0.001$. For the uncertainty parameters β, the rank-correlation between the generating and recovered parameters was $r_\tau = 0.52, p < 0.001$. For the greedy term γ, the rank-correlation was $r_\tau = 0.69, p < 0.001$. For the local search component λ, the rank-correlation between the generating and recovered parameters was $r_\tau = 0.75, p < 0.001$. For the novelty component ν, the rank-correlation between the generating and recovered parameters was $r_\tau = 0.52, p < 0.001$. For the large majority of participants (0.94, 67 out of 71), the softmax had a minimal value (exp(-5), set as minimum for numerical stability), indicating precise model predictions, and value-sensitive behaviour in participants. The values were slightly more dispersed in the recovered softmax parameters (19 simulations were fit with a softmax of exp(-5), the minimum possible value), and a median value of 0.02. the correlation was thus not very

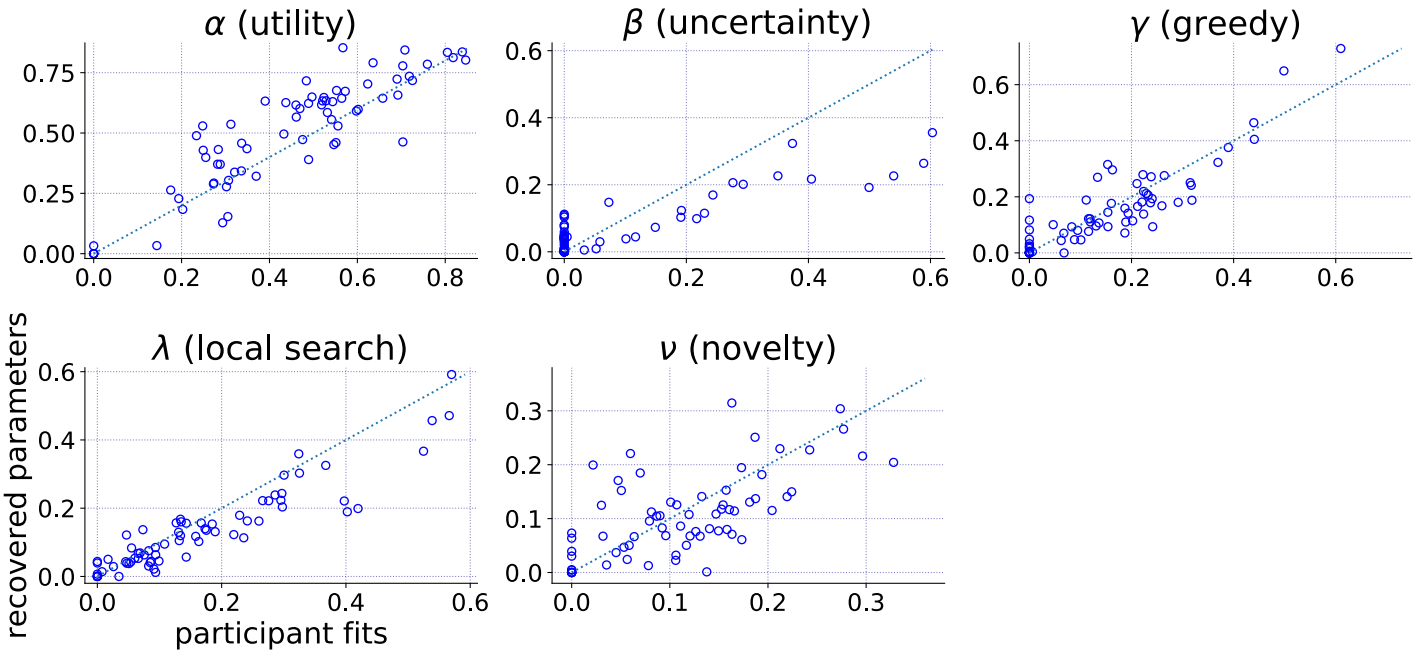


Figure 3.21: Recovered parameters from the model simulations generated from participant fits. *x*-axis displays the parameter value fit to participant data. *y*-axis displays the parameter value fit to the model simulation observations.

strong between generating and recovered parameters ($r_\tau = 0.18, p = 0.04$). This suggests that the model was more sensitive to the interaction between different independent components when fitting simulations than when fitting participant data.

As expected, the model log-likelihoods were significantly better when predicting model simulations ($M = -131.45, SD = 49.02$) than when predicting participant selections ($M = -155.7, SD = 40.36$) ($t(143) = -3.21, p = 0.002$). In summary, we find that the rank correlation between generating and recovered parameters was high across all parameters. This offers strong evidence in favour of the results of the general model to be reliable. The results may offer some evidence toward the existence of these different model components as independent psychological processes.

3.5.2 Qualitative analysis of model simulations

In this section we focus on some of the qualitative patterns observed in the empirical data, and study to what extent these can be found in our model simulations. To do this, we look at the behaviour of individual participants, and compare it to a model simulation generated from their set of parameters. We also compare the recovered parameters of the model simulation to the generating parameters in this case-by-case setting.

We first analyse one of the best performing participants, who traded off between exploration and exploitation and found the maximum value in two of the three grids, and the second highest tile in the first grid. The participant settled on a tile within approximately ten rounds and reselected it until the end of each grid (see Figure 3.22 a). In grid 1 and 2, this participant selected actions that were relatively distant from one another before doing minimal local exploration and

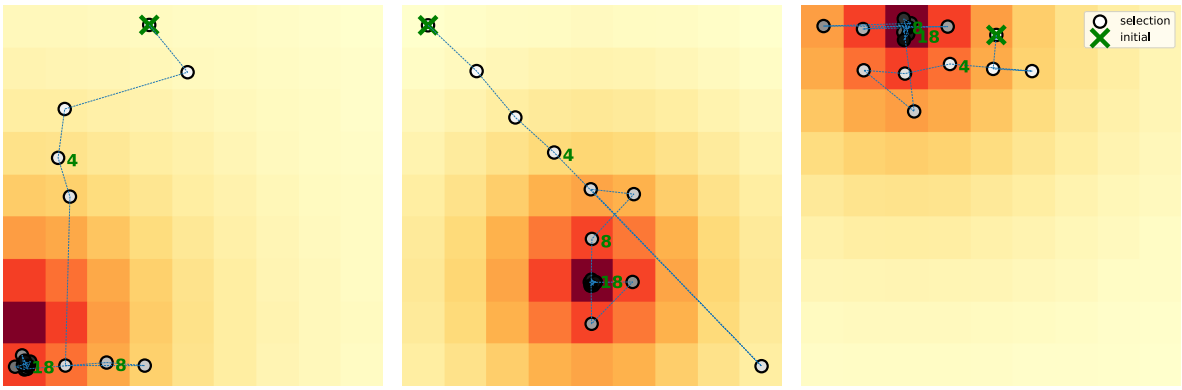
ending the search with greedy re-selection. In the third grid, the search process was exclusively local, possibly because the participant's initial selection was close to the maximum tile.

The model simulation (see Figure 3.22 b) captured these patterns fairly well, also finding the maximum value within ten observations, exploring with initial actions that were relatively distant from one another, and finishing the search with some local exploration steps. We plot the parameters fit to the participant data, and those recovered from the model simulation in Figure 3.23. In both cases, the most important component was the α parameter, or the expected mean under the GP model. Second was the greedy term. The local bias term and the novelty term both contributed slightly to the model predictions. The recovered parameters for the model simulation gave less weight to the novelty term than the generating parameters, but attributed some to the β term (uncertainty directed search), despite it not being given any weight in the participant data. Overall, we found the recovered parameters to be coherent with the parameters fitted to participants.

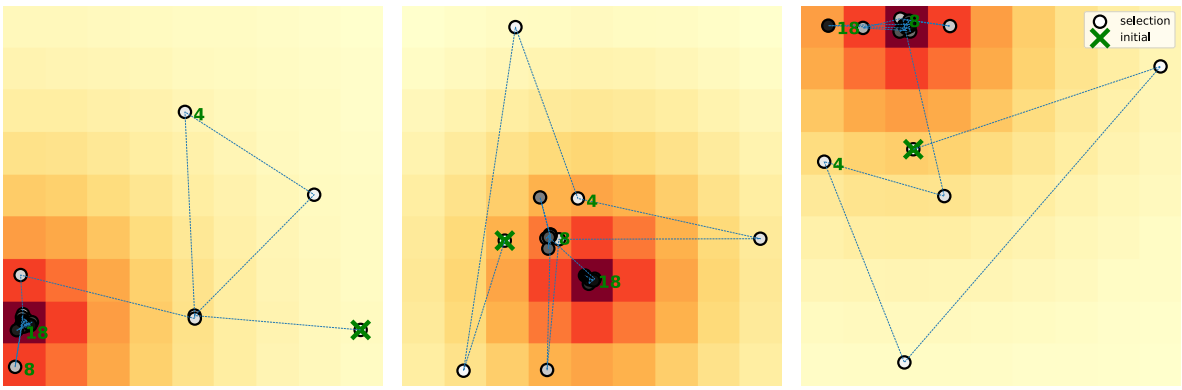
Next we look at one participant who engaged in *Full-Explore* behaviour (Figure 3.24). This participant preferred to choose tiles adjacent to their previous selections, with occasional longer jumps. The participant tended to select tiles near the maximum value tile, and found the maximum in every grid, but never re-selected it.

The model simulation was able to reproduce this local search behaviour with occasional jumps, and found the maximum in every grid without re-selecting it (except from in grid 3 where it was re-selected twice, see Figure 3.24).

As with the previous participant, the main driver was the expected utility term α, with the local bias and the novelty term being the other representative



(a) Participant selections



(b) Model simulation based on parameters fit to the participant.

Figure 3.22: Participant observations and model simulations across three grids. The parameters used by the model simulation were the ones fit to a participant. The green cross indicates the initial observation in each grid. The green number marks the trial at which some actions were selected. The relative tile value is marked through the colour of the tile. The circles show the observations, the darker a circle, the later in the round it was selected by the participant.

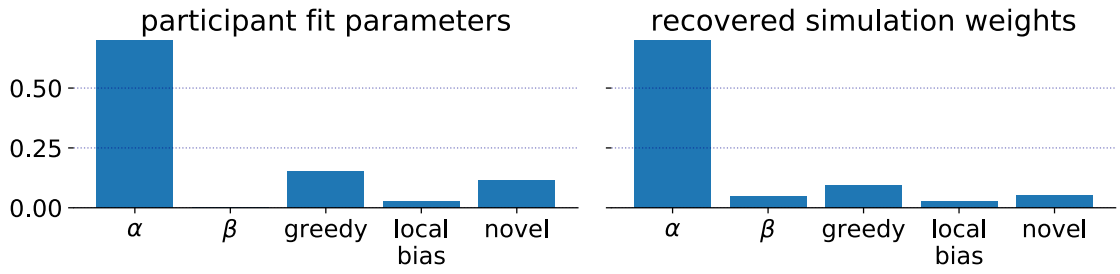
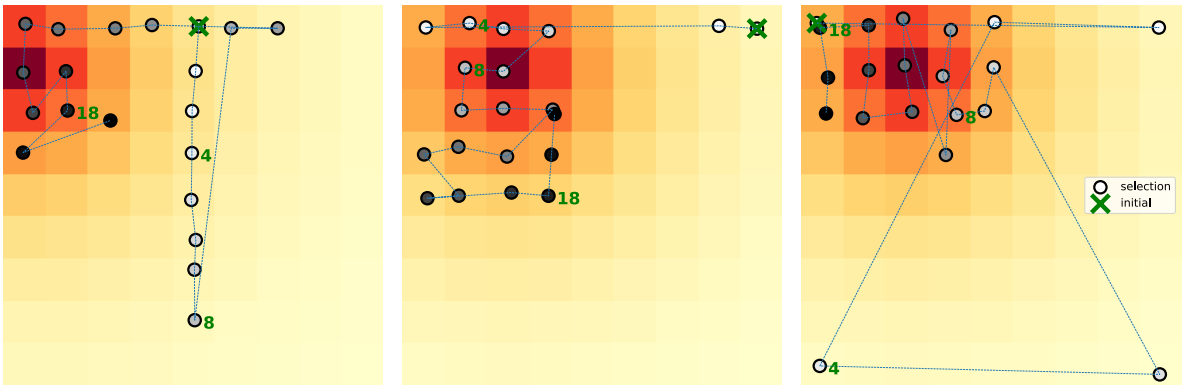
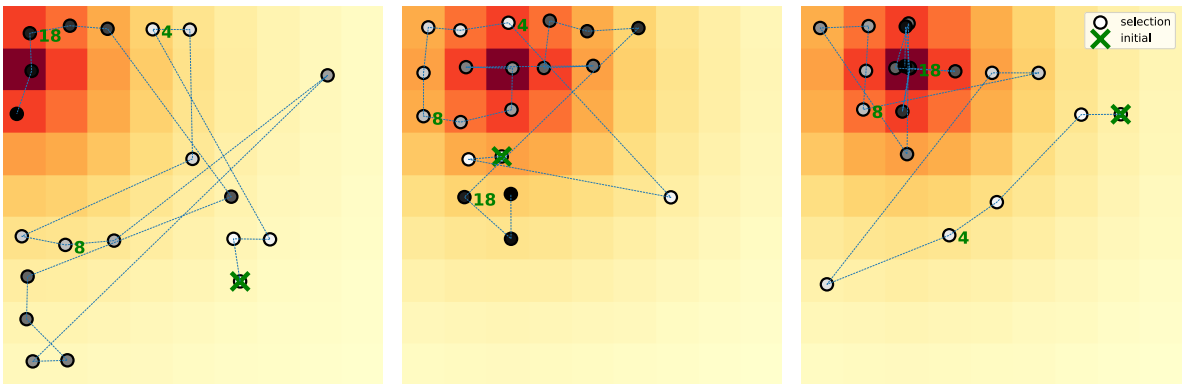


Figure 3.23: Comparison of parameters fit to participant and recovered parameters on model simulation.



(a) Participant selections



(b) Model simulation based on parameters fit to the participant.

Figure 3.24: Participant observations and model simulation across three grids. See Figure 3.22 for more details.

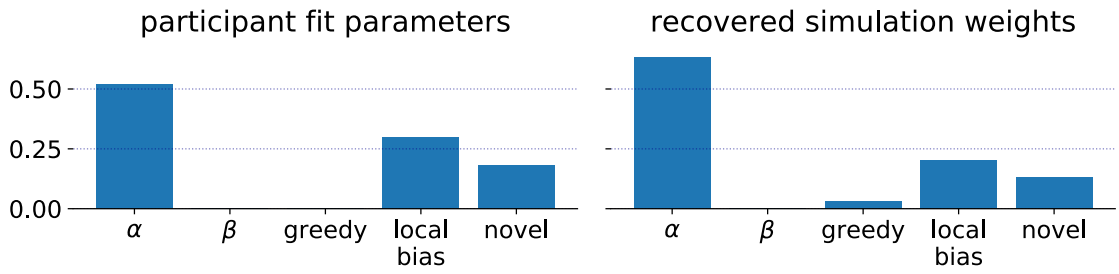


Figure 3.25: Comparison of parameters fit to participant and recovered parameters on model simulation.

components. This was largely recovered when fitting the model simulations, though the greedy term was given some minor weight, due to the re-selections in grid 3 (which were presumably driven by the α term, and the local search term λ).

Finally, we look at the performance of the model simulations and compare them to the performance of participants (see Figure 3.26). In Chapter 2, we found distinct patterns of performances amongst participants. Some participants performed consistently better than others, while many engaged in entirely exploratory behaviour and largely dismissed reward incentives. Here, we look at whether the model simulations based on the parameters fit to participants reproduce these distinct patterns of performance. We plot the performance of *Explore-Exploit* participants and *Full-Explore* participants separately, to highlight the differences between individual participants. We find that the general model is able to reproduce the characteristic patterns of performance for both sub-groups.

In summary, our general model can capture important qualitative patterns such as the amount of exploration, the type of exploration (global vs local), and performance, by modelling participants through a mixture of distinct search processes. In the next part we evaluate the predictive power of the general model and compare it to ablated versions of the model (i.e. by removing components)

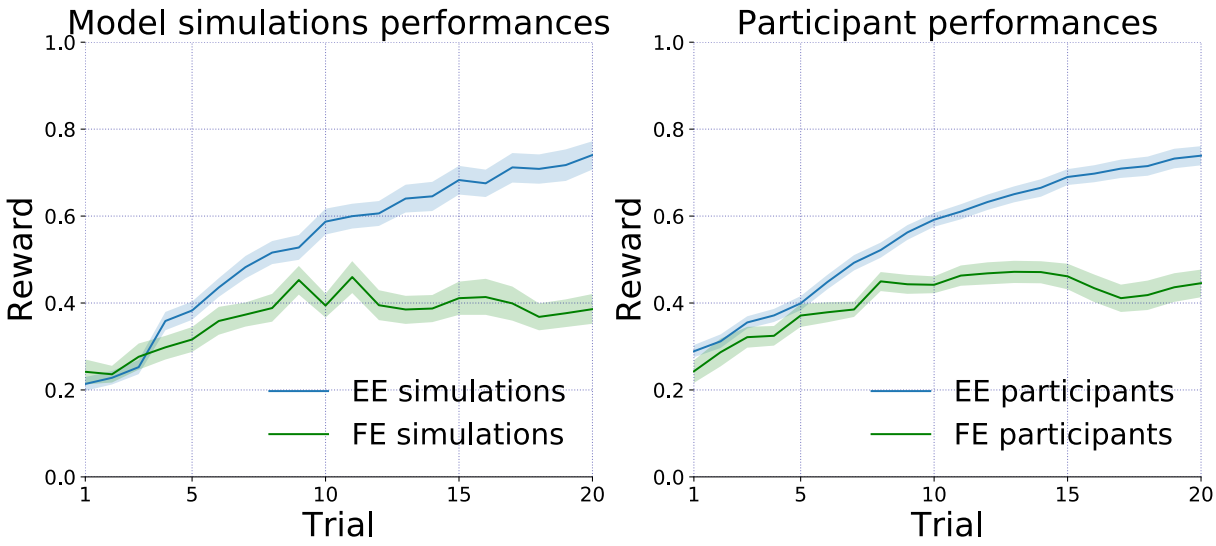


Figure 3.26: Performance of model simulations according to strategy type (*Full Explore* and *Explore-Exploit*) on left hand side. Right hand side shows corresponding participant performances.

to ensure they all contribute to capturing participant behaviour. In Section 3.6, we discuss some of the limits of the model, and qualitative patterns not captured through the simulations.

3.5.3 Model robustness and predictive power with comparison against ablated models

To assess the explanatory power of our general model, and the potential benefits of having a mixture of components, we fit it to the full set of participant observations and compare it to truncated versions of the model. The three truncated versions of the model we compare the general model to are:

- A GP-UCB only model with a free β parameter, fixed α (=1) with a softmax decision rule (two free parameters: β, τ).

- A GP-UCB with local-bias model i.e. a free β parameter, fixed $\alpha = 1$, a local search parameter and a softmax decision rule (three free parameters: β, λ, τ).

- A heuristics model, with the search, greedy re-selection and novelty components and softmax decision rule (four parameters: λ, γ, ν, τ).

We limited our comparisons to this subset of possible truncated versions of the model as they seem plausible as cognitive hypotheses. We also expected the resulting inferred parameters would inform us of the importance of the respective components given the L1 penalty imposed during the optimization.

We first compare the model likelihoods on the complete set of participant observations (i.e. all three grids). We use the Akaike Information Criterion (AIC) (Akaike, 1974) to take into account the different numbers of parameters across different models.

The general model provided the best complexity-penalised fit for 63 of 71 participants. A Mann-Whitney test shows that the general model (Mdn=318.57) was significantly better than the next best model (heuristics, Mdn=351.43) when comparing model AIC scores across participants (U(143)=1921.0, p=0.007). The model results point out that it was a combination of both heuristic strategies and generalisation-based strategies which carried explanatory power across individuals, as opposed to one or the other. This offers evidence to support that each independent component carries explanatory weight at the individual level, and not only across different participants. One risk of a highly expressive model is that it overfits participant data, by explaining the noise in participant decisions and failing to extract meaningful patterns, thus not yielding much predictive power.

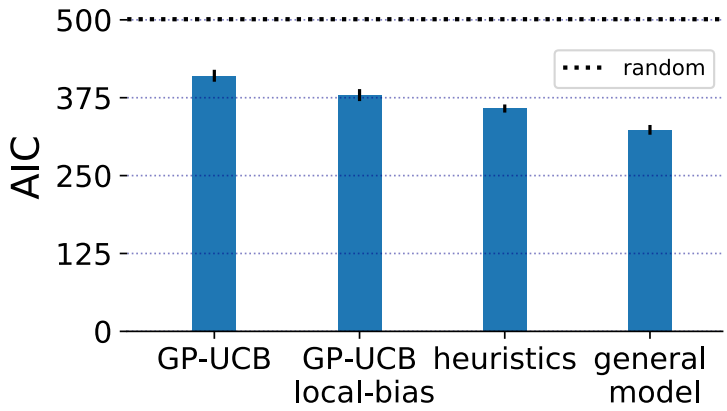


Figure 3.27: Mean AIC of general model and ablated model versions (lower is better). The error bar shows the SEM. The dotted line shows the AIC of a model predicting individual participants actions at random (uniform probability across all selections).

To evaluate this, we look at the accuracy of the general model against the simpler truncated versions of the model when predicting out-of-sample participant selections.

3.5.4 Model predictions on participant data

To ensure that the general model is not overfitting the participant data by adding unnecessary components and to evaluate the robustness of fits on unseen data, we examine the predictive accuracy of models on the selections made by participants in the third grid, with the parameters fit to the selections in the two initial grids. Like before, we compare the general model to the heuristic model, a GP-UCB baseline and a GP-UCB model with local-bias term.

Again the general model had the best predictive likelihood for the majority of participants (45 out of 71). 13 participants were best predicted by the heuristics

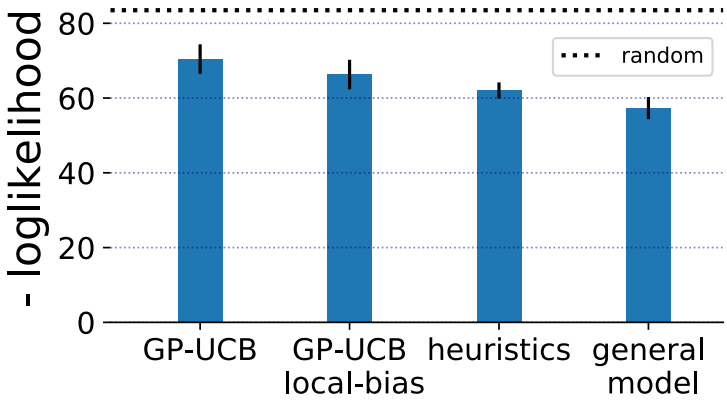


Figure 3.28: Predictive negative log-likelihoods of models on third grids (lower is better), based on the MLEs fit to participant observations in the first two grids. The error bars show the SEM.

model, 9 participants were best predicted by the GP-UCB local bias model, and 4 by the vanilla GP-UCB model. Because of the presence of outliers in the prediction scores, we use the Mann-Whitney test to compare the predictions of the general model against the next best model (the heuristics model). The general model was significantly better with a median predictive negative log-likelihood of 53.27, against a median of 61.83 for the heuristics model ($U = 1926.0, p = 0.007$).

We have shown the validity, robustness, and interpretability of our general model through several steps. We first recovered special case behaviours, to show that the fit parameters were distinct and coherent with the observed behaviour. Second, we ran simulations on the participant data (71 participants), and recovered the parameters from those simulations. The parameters were strongly correlated to the generating ones. We conducted a qualitative analysis of the model simulations and showed that the model was able to capture and generate qualitatively distinct types of search behaviour, as well as match the performance patterns of participants. Finally, we compared our general model to ablated versions of

the model and showed that participants were in general better explained by a rich combination of strategies, both individually and at a group level. This was also true for out of sample predictions, when the model was fit on the first two grids to make predictions about participant selections on the third.

In the next section, we focus on the limits of the model by examining some of the qualitative patterns observed in the data that were not captured by the general model and discuss how it could be further expanded.

3.6 Limits of the general model

In Chapter 2, we found that many participants were able to transfer knowledge across tasks. From one grid to the next, they were able to improve their performance. This was not the case for our model simulations, since we assumed in the model each task to be independent of the other (see Figure 3.29). We look at learning dynamics and how transfer can be explained in Chapter 6.

When looking at participants poorly predicted by the model, we find the model struggled to capture participants who had a sudden change in strategy. We show a participant who suddenly changed strategy in Figure 3.30. One possible explanation for this is that participants have a "toolbox" of strategies, and they adaptively selected based on their knowledge of the environment (Lieder & Griffiths, 2017). In this work, we focus on identifying and understanding these strategies. Understanding how people might switch between them is thus beyond the scope of this study.

In Section 3.4, we introduced a line-search heuristic inspired by patterns of behaviour observed in participants and used it as a special case model. We show three grids from three different participants in Figure 3.31. Although we

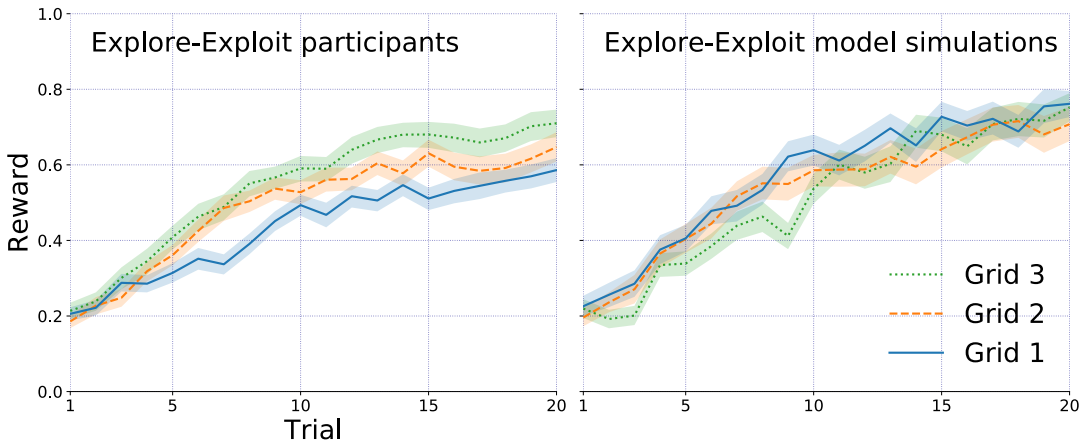


Figure 3.29: Transfer effects observed in Explore-Exploit participants, but not captured by the general model. Here the model simulations are run forward using the parameters fit to Explore-Exploit participants.

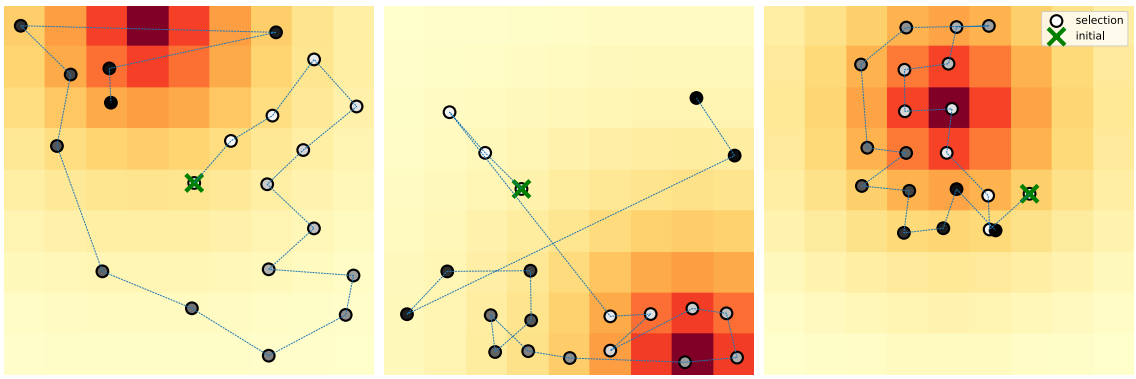


Figure 3.30: Example of a participant poorly predicted by model. The participant used global search in the first two grids, with clicks distant from each other. In the third grid however, they selected exclusively local actions, but started in a higher reward region.

were able to recover parameters that allowed us to broadly interpret this type of strategy, the model simulations were not able to reproduce the distinct patterns of behaviour seen in participants, characterised by local gradient-ascent steps in one dimension at a time. One of the benefits of such a strategy could also be to alleviate memory demands by only having to remember the direction, the previous tile reward and the maximum known tile, rather than the complete sequence of previous observations.

To recover this type of strategy, we implement a model that places a weight on the next tile in line based on the direction from the two previous observations (λ_1), on the tiles orthogonal to the previous selection (λ_2), and on the neighbours of the maximum known value (λ_3). These three terms define the exploration term. The model trades off between exploration and exploitation (re-selecting the maximum known value) explicitly through an ϵ parameter. The model predictions are then wrapped in a softmax function with a temperature parameter τ. We use the AIC to compare it to the general model and its ablated versions. We find that the line-search model best explains the behavior of 16 percent of participants (11 of 71). When looking at out of sample predictions on the third grid, this model best predicts the behavior of 25 percent of participants (18 of 71).

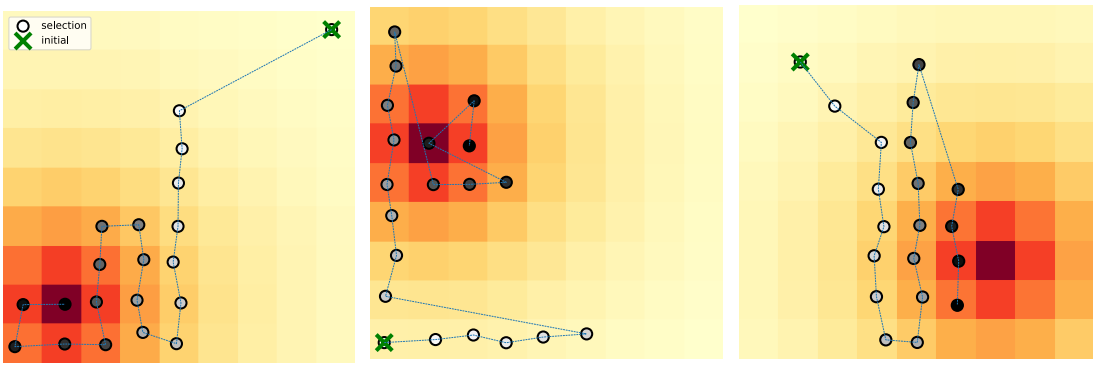


Figure 3.31: Examples of line search displayed by three different participants.

While the general model is unable to capture specific patterns of behaviour such as the line-search heuristic discussed here, it has the advantage of being able to

relate different types of strategies in a continuous psychological space. In the next section, we use the results from the general model to better understand the behaviour of participants across the different experiments.

3.7 Model based analysis of experimental results

In this part, we conduct a preliminary analysis of the results offered by the general model. In our analysis, we report model results on participants across all experiments (N=217). We exclude 11 participants from our analysis, as they were predicted at random, with all the components equal to zero.

When comparing the general model to ablated versions of the model, we found that the different components all carried explanatory power at a group level, but also at an individual level. Rather than being explained by a single component, participants were best explained as a mixture of the different search processes making up the model. The vast majority (.98) were best explained by a combination of 3 or more components, and the mode was a combination of four components (0.46) (see Figure 3.32). 0.96 of participants were best fit with non-zero values for the directed search (α), implying that people do rely on generalisation to guide their search. 0.91 with non-zero values for the novelty component (ν), and 0.89 with non-zero values for the local search component (λ). This suggests that participants had a strong inclination for novelty and local search, corroborating our empirical findings from Chapter 2.

The directed search component (α) was particularly important across all participants, with an average weight of 0.40 (SD=0.2), implying that generalisation and expected rewards explained participant selections. Conversely, the β parameter

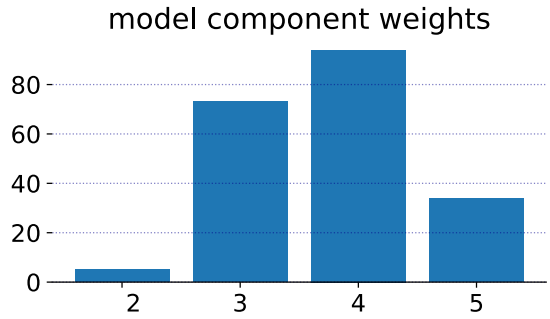


Figure 3.32: Distribution over the number of contributing components in the general model when fit to individual participants (N=206).

was the least important, with only 0.25 of participants being best explained with a non-zero value, meaning that uncertainty was not a big driver of exploration for participants. For all parameter distributions, see Figure 3.33.

In the next section, to better understand how changes in the environment affects participant strategies, we analyse the differences between conditions based on the model descriptions of participants.

3.7.1 The effect of data availability on participant strategies when learning across new tasks

We first compare the model results of participants in Experiment 1 and Experiment 2. In both experiments, participants were presented with three grids that followed the location rule. In Experiment 1 ($E1$), the rewards of a tile were shown for 1.5s after having been selected. In Experiment 2 ($E2_{vis}$), the rewards were continuously displayed once they had been selected. A brief recapitulation of the experiments presented in Chapter 2 is given in Table 3.1.

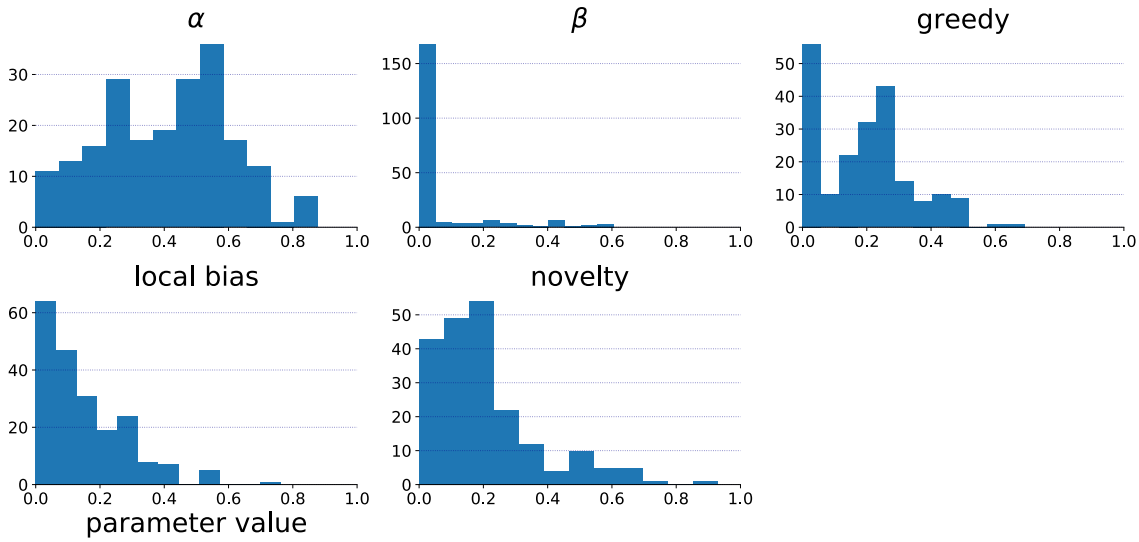


Figure 3.33: Distributions of the different model component weights across all participants (N=206).

Experiment	Training	Visible observations	N
$E1$	X	X	71
$E2_V$	X	✓	68
$E3_T$	✓	X	37
$E4_{TV}$	✓	✓	41

Table 3.1: The subscript E_T indicates the presence of training in the experiment, while the subscript E_V indicates that rewards remain visible on a tile once it has been selected.

Participants in Experiment 1 ($E1$) were modelled with the guided search component (α, Med=0.48) playing a significantly more important role than in Experiment 2 $E2_V$ (Med=0.33) ($U(140) = 1736.0, p = 0.002$). The local bias parameter (λ) was more important in $E1$ (Med=0.13) than $E2_V$ (Med=0.08) ($U(140) = 1916.0, p = 0.02$). The novelty term (or undirected exploration, ν) carried a much stronger explanatory weight in $E2_V$ participants (Med=0.21) than in $E1$ participants (M=0.11) ($U(140) = 1053.0, p < 0.001$). These results support the empirical analysis conducted in Chapter 2, namely that visible observations incentivised participants to explore more, to explore more globally and, as a result, were less concerned with maximising rewards.

3.7.2 The effect of data availability on participant strategies when learning on known tasks

To further understand the effect of visible observations on participant strategies, we look at Experiment 3 ($E3_T$) and Experiment 4 ($E4_{TV}$). In both experiments, participants were trained on the location rule before being presented with the three grids . In $E3_T$, like in Experiment 1, rewards disappeared after 1.5s. In $E4_{TV}$ rewards were constantly displayed after they had been observed (like in Experiment 2).

$E4_T$ participants' choices were driven by expected reward (α, Med=0.47), more so than $E3_{TV}$ participants (Med=0.35) ($U = 606, p = 0.06$). Six participants in $E3_{TV}$ had a non-zero uncertainty-directed weight (β, M=0.17, SD=0.13), while no participants in $E4_T$ were explained by it. Like for Experiment 1 and 2, there was a significant difference in the importance of the novelty bonus weights. It was significantly more important in $E3_{TV}$ participants (Med=0.24) than in $E4_T$ participants (Med=0.1) ($U = 288.0, p < 0.001$).

In $E1$ we saw that when the task structure was unknown and observations disappeared, participants had a significantly stronger tendency to explore locally than in $E2_V$ where observations remained visible. This effect was not observed when the task structure was known: both $E3_{TV}$ and $E4_T$ participants had a median value of 0.1 ($U = 694.0, p = 0.26$).

Overall, we found that both the availability of observations and training on the reward structure prior to the task both has important effect on participant strategies at a group level. We found that data availability led to more novelty driven exploration in participants, as opposed to more directed search when rewards disappeared. When participants were discovering the task structure, their exploration was more local when rewards disappeared than when they remained available. This was not the case when participants had been trained on the reward structure prior to the task.

3.8 Conclusion

In this chapter, we presented a general model to explain the strategies of participants that combines model-based processes as well as simpler heuristic strategies. The two model-based processes we considered in our model are *expected rewards* and *uncertainty*, which are both computed using a Gaussian Process model, a popular model of human generalisation. To account for some of the behaviours observed in participants in Chapter 2, we also considered simpler strategies that do not require a model of the world: Greedy re-selection of the best known action, random exploration and local search. Based on model simulations, we checked that the model was flexible enough to capture and distinguish a variety of behaviours, and the parameters used to describe participants were

interpretable. We also compared it to ablated versions and found it yielded better predictions than simpler models.

In general, we found that our general model was able to capture robust qualitative and quantitative patterns in how participants selected actions, and that most participants were best explained as a combination of these different processes as opposed to a subset of them. Our model offered strong evidence for people's ability to use generalisation to predict the value of unknown actions to guide their search. Local search, novelty and greedy re-selection of the max known were also strong predictors for people's behaviour. Surprisingly, there was little evidence in support of uncertainty driven search being an important driver for the actions of most participants.

We also found that there were consistent patterns in how participants selected actions, affected by the availability of information when they explored, and their knowledge of the reward structure. The general model allowed us to explain the influence of the different experimental controls on the search strategies used by participants. Finally, we looked at some of the qualitative patterns that the model was not able to capture, and discussed how the model could be further improved. In this chapter, we limited our analysis to group level differences. In Chapter 2, we highlighted that beyond the group level differences, important differences at the individual level also existed. In the next chapter, we develop a model of individual differences using the results of the general model to better understand the similarities and differences across participant strategies.

Chapter 4

Understanding similarity and differences in human strategies

4.1 Introduction

So far, we have discussed the results of four experiments through an empirical analysis of participant behaviour. Across all four experiments, participants were presented with a task of similar structure, yet we observed clear differences in their strategies. In the last chapter, we introduced a general computational framework to study the behaviours of participants. When motivating our general model, we discussed the benefits of using model expansion to capture competing hypotheses as special cases of a general model. One of the benefits we highlighted was the ability to capture a diverse set of behaviour within a continuous parameter space. In this chapter we focus specifically on the problem of understanding the differences between participants by leveraging this shared psychological space.

The diversity of strategies used by people in explore-exploit tasks, and the study

of individual differences, has been a subject of interest for a number of studies. Steyvers *et al.* (2009), for example, studied the decisions of 451 participants on bandit problems and found clear evidence for individual differences, and reported correlations between participant decisions and a set of psychological variables. In their study, individual differences are explained through different heuristics best describing individuals. Yi *et al.* (2009) looked at individual differences in a restless bandit task, and reported substantial variation in the overall performance and the degree to which participants switched between options. Here, the difference in participant behaviour was explained as the result of different parametrisations of a particle filter algorithm. Similarly, Reverdy *et al.* (2014) classified 326 participants according to three distinct models of regret on a spatially correlated multi-armed bandit task. In a more applied domain, understanding the variation in novelty seeking behaviours is of central interest to clinical research. Indeed, a growing number of studies have started to look at how systematic differences in novelty seeking strategies across individuals to make sense of behaviours that relate to addiction, impulsive behaviour or risk-taking (Djamshidian *et al.*, 2011; Addicott *et al.*, 2013, 2017; Harlé *et al.*, 2015, 2017; Clark *et al.*, 2013).

In Chapter 3, we presented a general model to capture human exploratory strategies. We showed that different sets of decision strategies could be captured and represented by a unique parametrisation under this general model. The approach of modelling individuals independetly has the benefit of not losing information or corrupting the data by aggregating or averaging it. However, it also carries downsides. Human decisions are inherently noisy and difficult to model. Conducting group level analyses can remove some of this noise and help overcome the sparsity of the data of a single individual. Only focusing on an analysis at the individual level carries the risk of overfitting by modelling the

noisy elements of participant decisions, making it more difficult to extract the more general patterns of behaviour in the data and generalise to other contexts.

In this chapter, we attempt to overcome these issues by uncovering "families" of strategies shared amongst participants. To do this, we model participants under the assumption that each belongs to one of potentially many groups of possible strategies. Within a given group, participants will behave in a similar way, but there can be a number of different groups and participants will behave differently from one group to the next. In other words, even if each participant is unique, the variations between participants are not random. In this modelling framework, the groups observed in the our data set are not understood as a fixed set of strategy types that fully explains the variation between participants. Instead, as components that are a part of an arbitrarily rich structure. Given more data, richer and finer grained details about individual differences can be uncovered as the number of inferred groups grows (Navarro *et al.*, 2006). This intuition can be modelled in a hierarchical Bayesian framework, where the differences between participants are described with a distribution over participant parameters. We thus have a model at the level of the individual, that describes the selections of participants with parameters $\boldsymbol{\theta}$, where θ_i is the parameter corresponding to participant i, and a model at the group level that describes the differences between participants with parameters $\boldsymbol{\phi}$, where ϕ_j corresponds to the parametrisation of one of the groups j.

4.2 Evaluating the diversity of participant strategies

To gain a first intuition of the diversity of participant strategies as captured by the general model we transpose the 6D parameter space into 2D using the

T-distributed Stochastic Neighbor Embedding (t-SNE) algorithm (Maaten & Hinton, 2008). This is shown in Figure 4.1. t-SNE constructs embeddings by minimising the Kullback-Leibler divergence between the joint probabilities of the low-dimensional (2D) embedding and the high-dimensional data (6D). One of the advantages of t-SNE is its ability to capture the local structure of the data (close points are embedded close to each other) while also revealing some important global structure (such as clusters). We were interested in seeing if clear structures would emerge from the embedding, thus corroborating the approach of modelling individual differences as stemming from groups of strategies, using a hierarchical modelling approach. We were also curious to observe whether participants from the same experimental conditions would be clustered together. While the 2D representation should not be used to make strong conclusions, it can give a first intuition about the existence of patterns in the data.

On inspection, some clusters seem to exist in the data, but the embedding does not suggest that these clusters are directly determined by the experimental conditions of participants. This may suggest that the types of strategies used by participants go beyond the features of the tasks they were presented with (i.e. training vs no training, visible vs disappearing rewards). This could potentially point at different processes at play underlying the mechanisms behind human exploration. Clusters at the top-left and on the right hand side of the map (mostly $E2_V$ and $E1$ participants) do however suggest that similarities in experimental context (no training, or visible observations) might bear some influence on the strategies participants used. It seems that patterns in variation of participant strategies exist both at the individual and at the group level, which further supports the idea that studying individual differences can be informative to better understand both the changing and the invariant aspects of people's decisions strategies. In the next part, we attempt to capture these patterns of variations by constructing a model of individual differences.

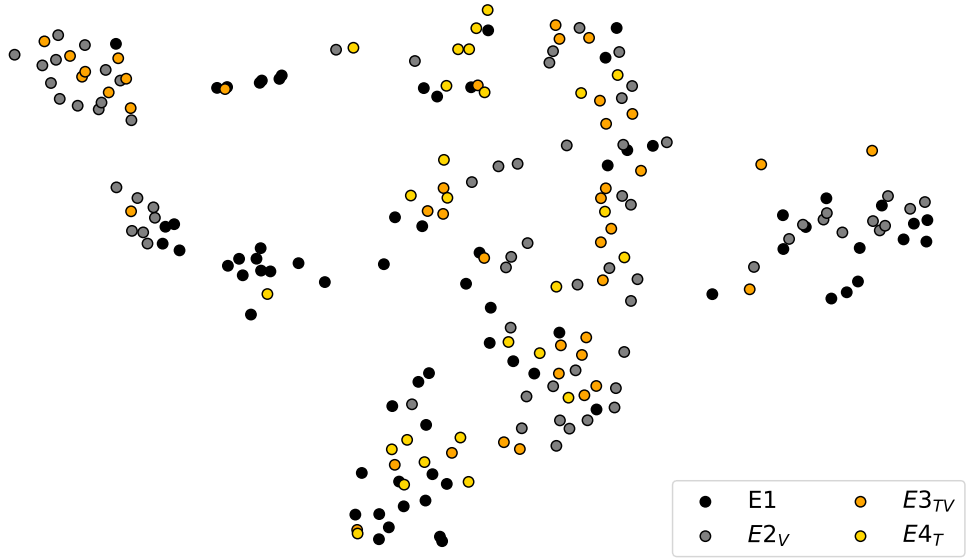


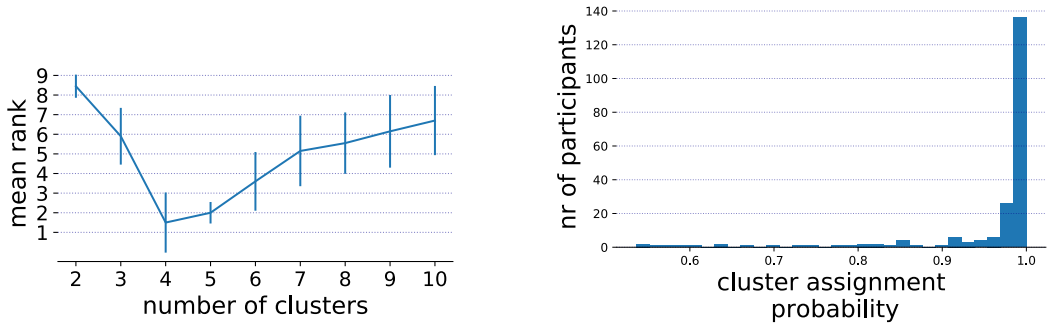
Figure 4.1: Map over different the strategies used by participants, as captured by the general model. Visualisation shows t-SNE embedding of parameters fit to participants, and their respective experimental condition. In E1 and E2, participants presented with unknown structures, while E3 and E4 had prior training on the nature of the task structures.

4.3 Identifying clusters of strategies

To identify the different groups of participant strategies, we use a Gaussian Mixture Model (GMM) on the 6D parameter space from the general model fits obtained through MLE. The GMM is an unsupervised clustering algorithm that assumes the data to be generated by K clusters, each associated with a parameter θ_k. Each cluster corresponds to a group of strategy. We estimate the number of clusters using leave-one-out cross validation. Participants were split into 10 groups (folds) at random and we fit the GMM model on all participant folds except for one. The mean negative log likelihood was estimated for a range of groups K=1,..., 10 in each cross-validation fold. 11 participants were predicted at random by the model (i.e. all weight parameters were zero) indicating random/unpredictable behaviour. We excluded them from the participant data set assuming that they corresponded to a separate cluster. On average, the cross-validation of the GMM indicated that a separation into four subgroups provided the most generalisable model, meaning that it yielded the lowest negative log-likelihood score on the test set (see Figure 4.2). We also evaluated the predictive power of different clusters by using the BIC (Schwarz *et al.*, 1978) and silhouette score (Rousseeuw, 1987), which measures group cohesion and distance to other clusters. Increasing the number of clusters above 4 did not yield significant improvements to the BIC. The silhouette score, also strongly favoured 4 clusters.

When looking at the cluster assignment probabilities, 95% of participants were assigned to a cluster with a probability larger than 0.7. The cohesion of clusters supports the hypothesis that distinct families of strategies amongst participants exist.

Based on the parameters of the cluster centres, we give the different groups the following names: *Greedy local* (n=47), *Scholars* (n=35), *Local explorers* (n=48) and *Maximisers* (n=76). We explain our nomenclature and the different



(a) Average rank of GMM with different number of clusters when comparing scores on test set using 10-fold cross validation. The results shown are from 20 evaluations.

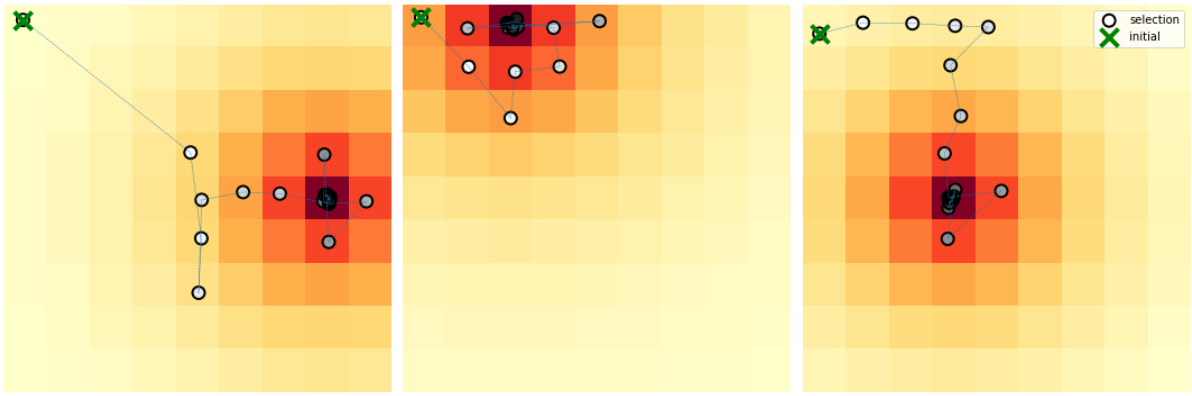
(b) Cluster assignment probabilities across participants for K=4 clusters. 95% of participants were assigned to a cluster with a probability larger than 0.7

Figure 4.2: GMM cross-validation results and cluster assignment probabilities with K=4 clusters.

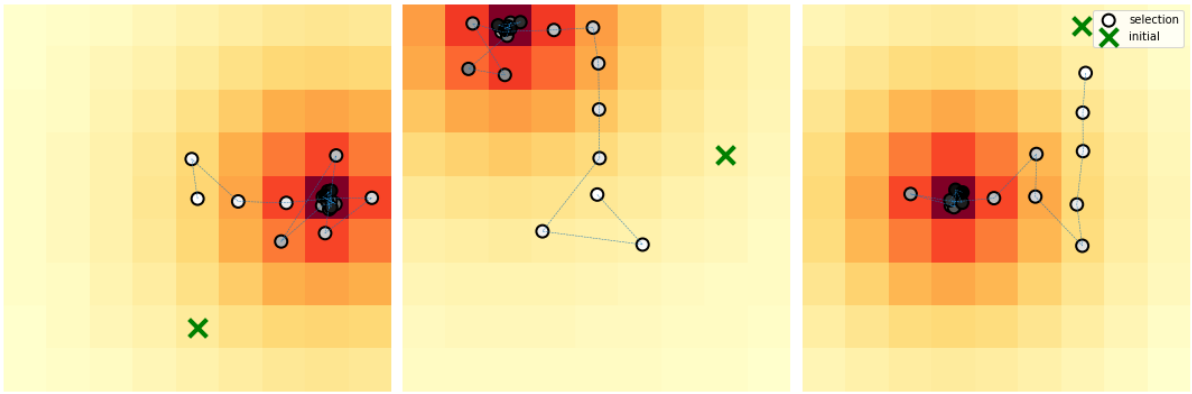
group characteristics below. The parameters of each cluster centre are shown in Figure 4.3. The largest group, the *Maximisers*, described participants best clustered under a dominant GP expectation term (α=0.56). The expected rewards under the Gaussian Process model denote an ability to generalise based on similarity to previous observations. This behaviour was the most frequent across all conditions all except for in $E2_V$. We show two participants from the *Maximiser* cluster in 4.4. In the two examples, we find that participants explore mainly through small steps, incrementally ascending towards the maximum and then efficiently re-selecting it without any further exploration.

Participants in the *Scholar* cluster have a cluster centre with two important GP components: Expected rewards and uncertainty driven search. It is the only cluster centre with a non-zero β term (β=0.30). We show two example participants from the Scholar cluster in Figure 4.5. Here, the two participants engage in initial exploratory actions that are very far from one another and eventually settle on the maximum tile or one of its neighbours. For both Scholars and Maximisers the most important components are the GP-driven components,

Figure 4.3: Cluster centre parameters obtained from the GMM clustering algorithm.



(a) Maximiser participant in Experiment 1



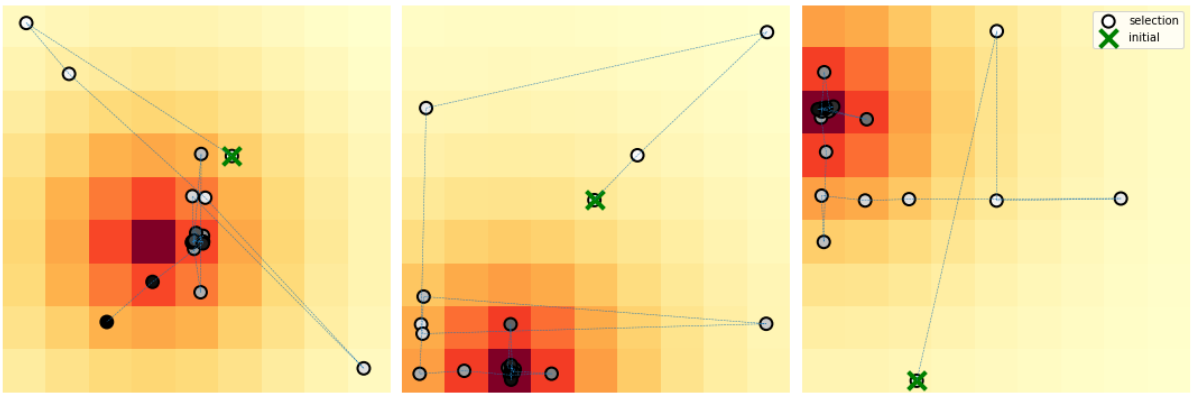
(b) Maximiser participant in Experiment 2

Figure 4.4: Two participants under the *Maximiser* cluster showing qualitative similarities in strategy despite the differences in experimental condition.

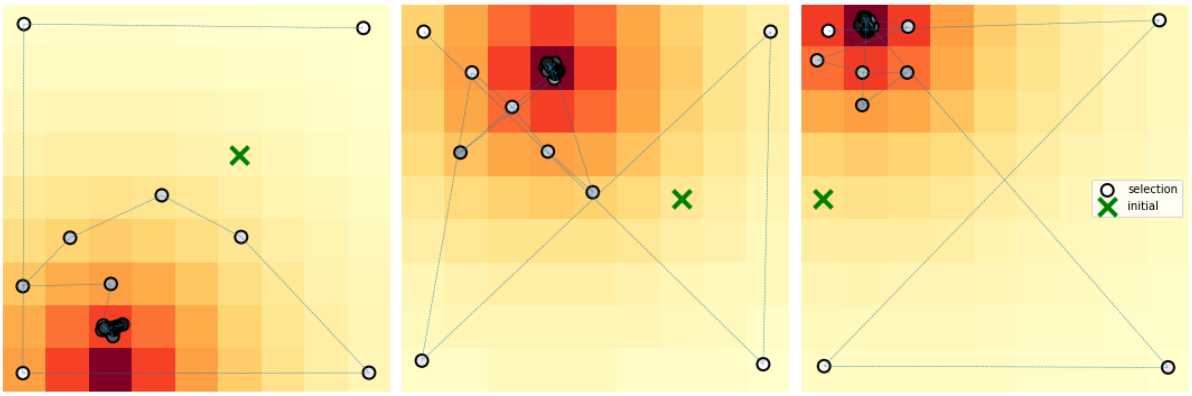
indicating model based strategies and supporting people's use of generalisation to guide their search. The *Local explorer* participants are clustered around a centre with novelty as its most important component, combined with expected rewards and a strong local bias term. It is the only cluster centre with no weight on the greedy component. Combined with an important novelty term, it suggests a *Full explore* strategy discussed in Chapter 2. We show two participants, from Experiment 2 and 3, belonging to this cluster in Figure 4.6. Here, the participants engage exclusively in exploration never reselecting the same tile. They mainly explore locally, selecting direct neighbours of their previous selection, with some occasional longer jumps. We name the last group *Greedy local*, since its centre's most important parameter is the greedy component while also having an important local bias term. The *Greedy local* cluster centre shares the same "active" model components as the *Maximiser*, with a more important greedy term and less emphasis on the expected rewards. We show two participants from the Greedy local cluster in Figure 4.7. The two participants engage in a mix of global and local exploratory actions and stop early, re-selecting a sub-optimal tile.

Across these participant examples, we note first strong qualitative differences between the different strategy types presented. Conversely, we also find obvious similarities between participants belonging to the same clusters, even when they came from different experimental conditions, thus justifying collapsing the data across our four experiments to study participant strategies.

We show the t-SNE embedding and mark participants to their respective clusters in Figure 4.8. We find that the clusters obtained by the GMM correspond neatly with the structure of the t-SNE embedding. To better understand the patterns of behaviour of participants, we use the clusters as tool of analysis in the next section.

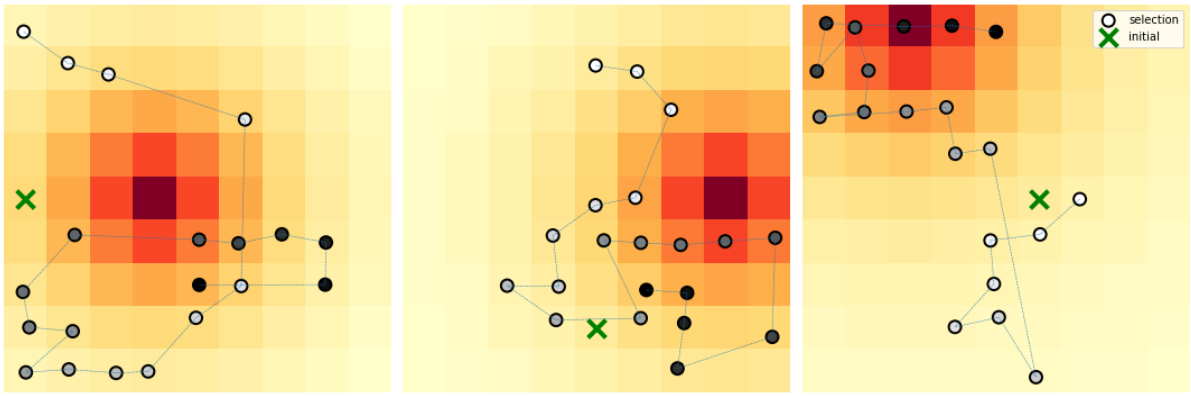


(a) Scholar participant in Experiment 1

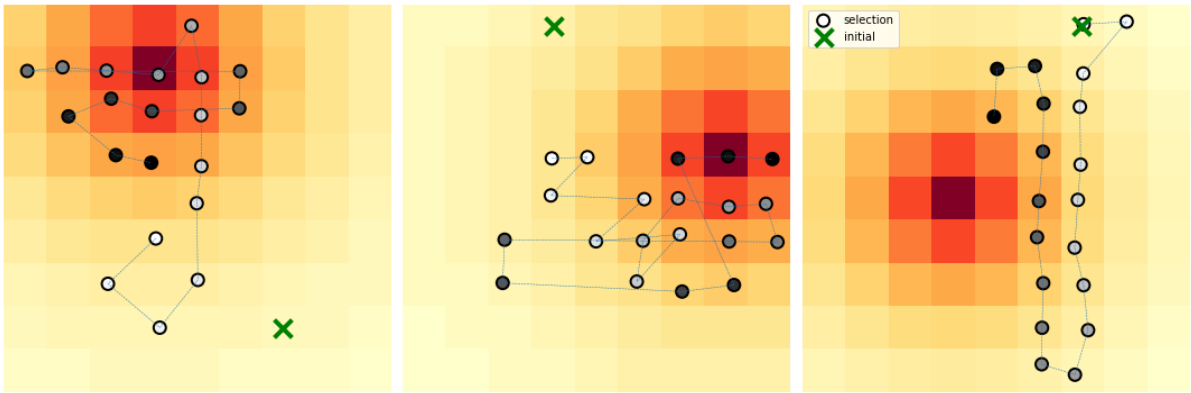


(b) Scholar participant in Experiment 2

Figure 4.5: Two participants under the *Scholar* cluster showing qualitative similarities in strategy despite the differences in experimental condition.

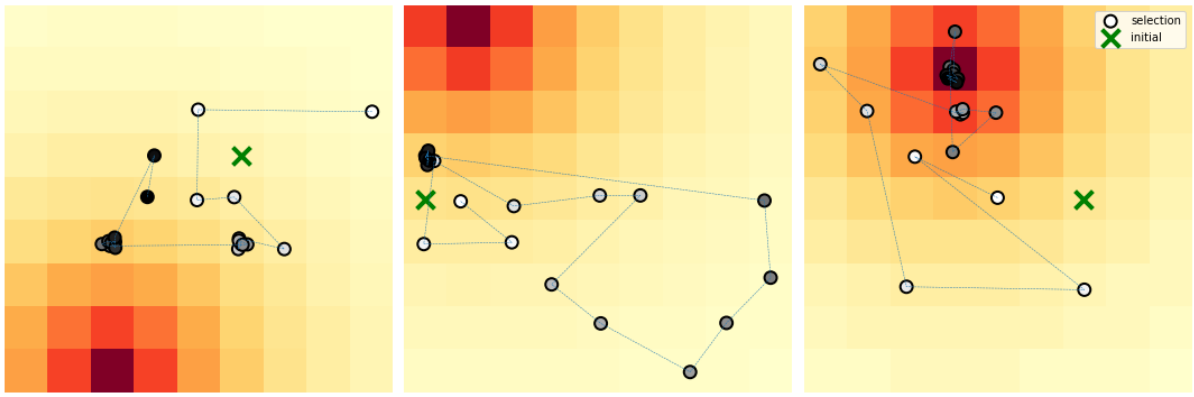


(a) Local explorer participant in Experiment 2

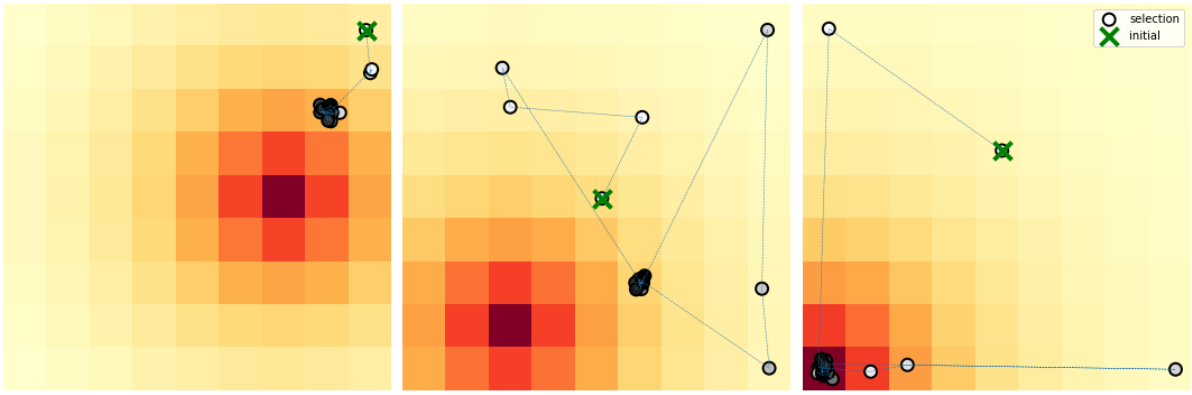


(b) Local explorer participant in Experiment 3

Figure 4.6: Two participants under the *Local explorer* cluster showing qualitative similarities in strategy despite the differences in experimental condition.



(a) Greedy local participant in Experiment 2



(b) Greedy local participant in Experiment 4

Figure 4.7: Two participants under the *Greedy cluster* cluster showing qualitative similarities in strategy despite the differences in experimental condition.

Figure 4.8: Map over different the strategies used by participants, as captured by the general model. Visualisation shows t-SNE embedding of parameters fit to participants, along with GMM clustering (on the 6D data).

4.4 Experimental analysis using strategy types

4.4.1 Choice of strategy given environmental features

Figure 4.13 shows the ratio of the different subgroups across all four experiments. $E1$ had a fairly uniform distribution of subgroups, with 0.24 of greedy local participants, 0.27 as Scholars, 0.19 as Local explorers and 0.30 as Maximizers. In $E2_V$, there were significantly more Local explorers than in E1 ($p = 0.05$). There were significantly fewer Scholars in $E3_{TV}$ when compared to $E1$ ($p = 0.05$). In $E4_T$ there were no Scholars and a single local explorer. In general, the visible observations feature had a strong effect on the ratio of Local explorers, whereas training had an effect on the number of Scholars.

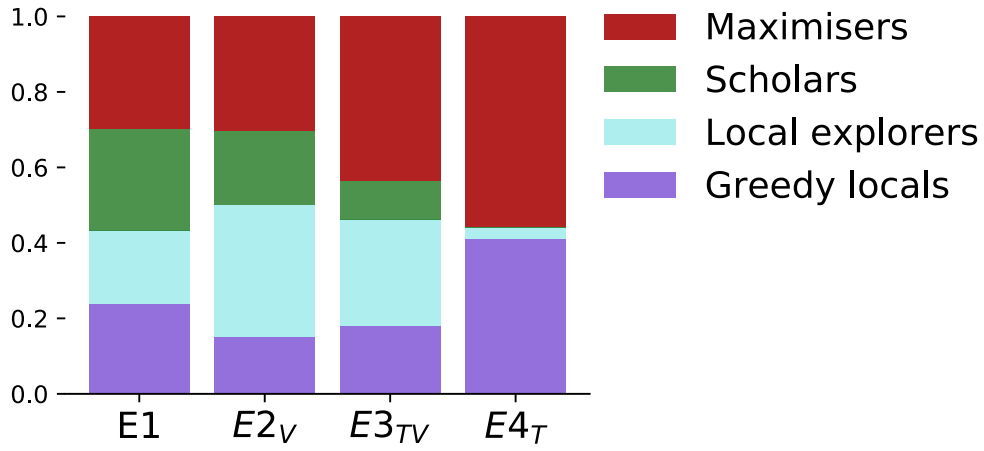


Figure 4.9: Ratio of subgroups across experiments

Across all experiments, the ratio of participants participants predominantly described by model-based components (i.e. Scholars and Maximisers) and of participants mainly described by heuristic strategies (Greedy locals and Local explorers) was fairly constant (in same order, M=0.57, 0.5, 0.54, 0.56). One possible explanation is that experimental manipulations (i.e. training and

information availability) did not influence participants' inclination to engage in model-based exploration, but influenced more specifically the type of strategy when engaging in either model-based or model-free behaviour. The determinant of whether participants engaged in model based strategies could have been down to the individual participant, e.g. their assumptions about the complexity of the task, or how much cognitive effort they put in. This hypothesis could be tested e.g. by looking at how cognitive load of time pressure affects participant strategies, under the assumption that heuristics strategies are cheaper than model based ones, and would thus be preferred.

Next, we look at the performance of the different sub-groups across the four experiments. In general, we hypothesised that the performance of participants would be distinct across the four subgroups of participant strategies, and that there would be patterns corresponding to their model descriptions. One prediction was that Scholars and Maximisers would display transfer effects in Experiment 1 and 2, and that they would outperform Greedy locals and Local explorers across all four experiments, under the assumption that leveraging a model of the environment would yield substantial performance benefits.

4.4.2 Experiment 1

We first compared the performance of the different groups by 1-way ANOVA test on the average performance of participants across all three grids. This showed that there were significant differences in the average performance of the different groups. A Tukey HSD post-hoc comparison showed that Maximisers significantly outperformed all the other groups. Scholars and Greedy locals both outperformed Local explorers, but the difference between the two groups was not significant ($\Delta_M = 0.04$). We then looked at patterns in how participants progressed within

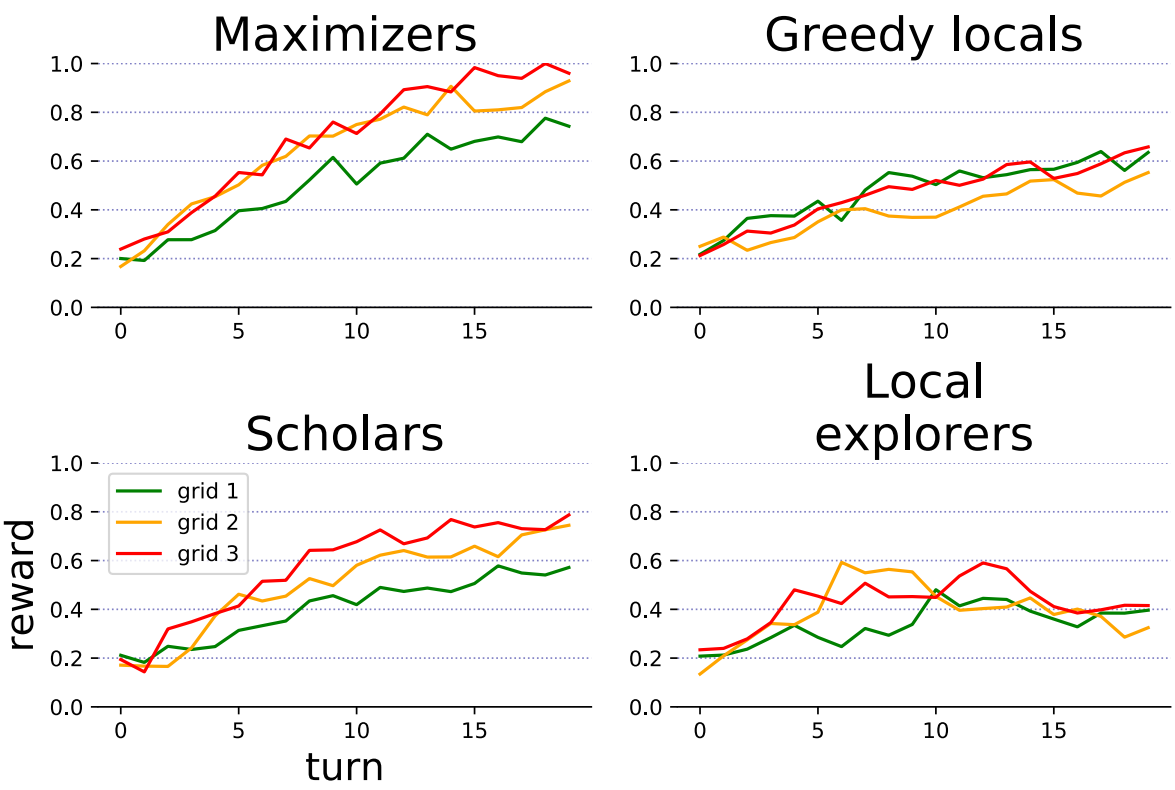


Figure 4.10: Performance of subgroups in Experiment 1 (E1, N=67). Greedy locals: n=16 participants, Scholars: n=18, Local explorers: n=13, Maximisers: n=20.

grids and across all three grids. Participants in the Maximiser (n=20) and Scholar (n=18) groups demonstrated progress across trials (Maximisers: $b = 0.04, se = 0.001, p < 0.001$; Scholars: $b = 0.03, se = 0.001, p < 0.001$) and across grids (Maximisers: $b = 0.09, se = 0.008, p < 0.001$; Scholars:$b = 0.08, se = 0.009, p < 0.001$), showing that they were able to learn the underlying task structure and exploit it efficiently (see Figure 4.10). Greedy local participants (n=16) did not show any progress across grids ($b = -0.01, se = 0.01, p = 0.47$), but were able to increase their score with the number of trials ($b = 0.02, se = 0.001, p < 0.001$). Conversely, participants within the Local explorer cluster (n=13) had very little progress across trials ($b = 0.01, se = 0.001, p < 0.001$), but some slight progress across grids ($b = 0.04, se = 0.01, p < 0.001$).

Next, we looked at the distance between exploratory selections of participants. A 1-way ANOVA test showed that the means were significantly different across the different groups ($F = 14.02, p < 0.001$). A Tukey HSD post-hoc comparison showed that Scholar participants had significantly longer average distances between clicks (M=3.06, Local explorers: M=2.1; Maximisers: M=1.9; Greedy locals: M=2.4) than in the other three groups, supporting the theory that these participants engaged global exploration to reduce structural uncertainty. The were no significant differences amongst the other three groups.

We conducted the same analysis when looking at the ratio of exploration participants engaged in (where exploration is defined as selecting an unseen tile). Again, the means were significantly different ($F = 15.91$, $p < 0.001$). A Tukey HSD post-hoc comparison showed that Local explorers had a significantly higher exploration ratio (M=0.96) than the other three groups of participants (Scholars: M=0.63; Maximisers: M=0.56; Greedy locals: M=0.67). The differences between the other groups were not significant.

4.4.3 Experiment 2

In Experiment 2, the average performances were also significantly different across the different groups ($F = 36.47, p < 0.001$). A Tukey HSD post-hoc comparison showed significant differences between all groups, except for between Greedy locals (n=10) and Local explorers (n=23). Maximisers (n=20; M=0.62) were the best performing participants, followed by Scholars (n=13; M=0.52), Greedy locals (M=0.38) and Local explorers (M=0.34). Participants under the Scholar cluster followed a similar pattern as in Experiment 1. They demonstrated progress across trials ($b = 0.03, se = 0.002, p < 0.001$) and across grids ($b = 0.06, se = 0.11, p < 0.001$). Maximisers showed the same progress across trials (b=0.03, se=0.001, p¡0.001), but only slight progress across grids

($b = 0.03, se = 0.009, p < 0.001$). Local explorers selected only slightly more rewarding tiles across trials ($b = 0.01, se = 0.001, p < 0.001$), and no transfer across grids ($b = 0.006, se = 0.007, p = .79$). Greedy locals showed weak progress across trials ($b = 0.01, se = 0.002, p < 0.001$) and weak transfer across grids ($b = 0.03, se = 0.01, p = 0.03$).

When looking at distance across selections, only Scholars had significantly longer distances between clicks than all other three groups ($F = 15.13, p < 0.001$). The amount of exploration in each group marked stronger differences across the different groups ($F = 42.41, p < 0.001$). Like in Experiment 1, Local explorers explored significantly more than all other three groups (M=0.99). Greedy locals (M=0.74) also had a ratio of exploration significantly higher than Maximisers and Scholars. There was no significant difference between Maximisers (M=0.54) and Scholars (M=0.54).

4.4.4 Experiment 3

Performances in Experiment 3 showed significant differences amongst participants clusters (F=15.05, $p < 0.001$). Again, Maximisers were the best performing participants (n=17; M=0.69), followed by Scholars (n=4; M=0.57), Greedy locals (n=7; M=0.47) and Local explorers (n=11; M=0.36). A Tukey HSD post-hoc comparison showed the differences were significant between Maximisers and Greedy locals and Local explorers. Scholars also significantly outperformed Local explorers. The average performance was directly related to the groups learning rates across trials: Maximisers had the fastest progress across trials ($b = 0.04, se = 0.001, p < 0.001$), followed by Scholars ($b = 0.02, se = 0.003, p < 0.001$), Greedy locals ($b = 0.01, se = 0.002, p < 0.001$) and Local explorers ($b = 0.003, se = 0.001, p < 0.041$). Because of the training prior to the tasks, there was no evidence for transfer across grids in the different groups. Like in

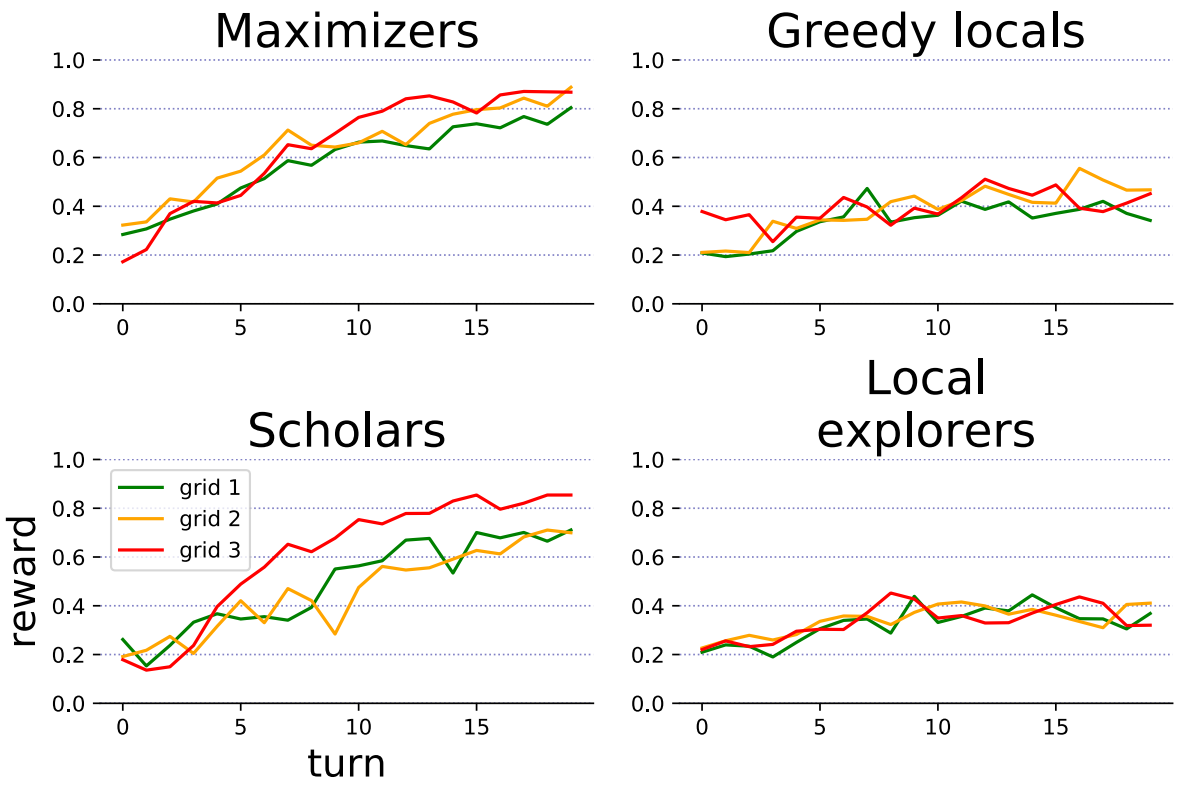


Figure 4.11: Performance of subgroups in Experiment 2 ($E2_V$, N=66). Greedy locals: n=10 participants, Scholars: n=13, Local explorers: n=23, Maximisers: n=20

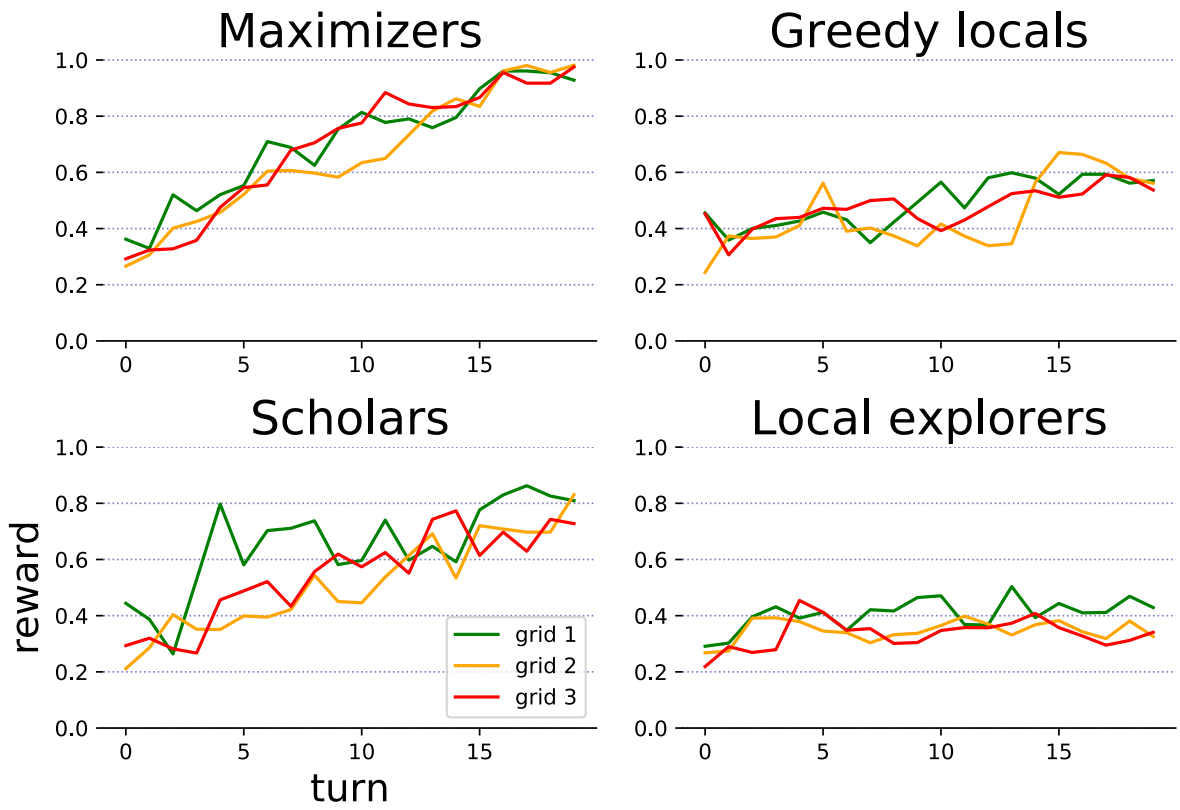


Figure 4.12: Performance of subgroups in Experiment 3 ($E3_{TV}$, N=39). Greedy locals: n=7 participants, Scholars: n=4, Local explorers: n=11, Maximisers: n=17

Experiment 1 and 2, Scholars had significantly longer distance between their selections than in all other groups ($F = 7.80, p < 0.001$). There was no significant differences between the other cluster of participants. Similarly, just like in Experiment 1 and 2, Local explorers had significantly more exploratory selections (M=1.0) than Scholars (M=0.64), Maximisers (M=0.6) or Greedy locals (M=0.61).

4.4.5 Experiment 4

In Experiment 4, there was only one participant clustered under the Local explorer group, and none in the group of Scholars. We thus limit our analysis to Greedy

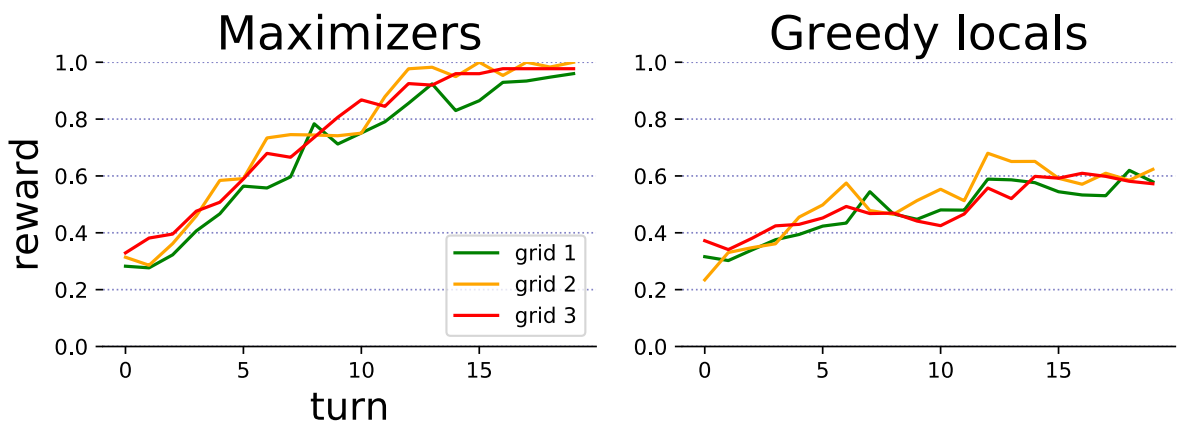


Figure 4.13: Performance of subgroups in Experiment 4 ($E4_T$). Greedy locals: n=14 participants, Maximisers: n=19.

locals and Maximisers. The average performance of Maximisers (n=19) was significantly better than Greedy locals (n=14) (t=6.92, $p < 0.001$). Greedy locals progressed slightly across trials ($b = 0.002, se = 0.002, p < 0.001$), but not across grids ($b = 0.001, se = 0.01, p = 0.61$). Conversely, Maximisers showed significant progress across trials ($b = 0.04, se = 0.001, p < 0.001$). Despite training, they also showed slight improvement across grids ($b = 0.03, se = 0.007, p < 0.001$). There was no significant difference in the distance between selections of Maximisers (M=1.90) and Greedy locals (M=2.0) ($t = -0.36, p = 0.73$), nor between their exploration ration ($M = 0.52, 0.52; t = -0.07, p = 0.95$).

4.5 Conclusion

In this chapter, used a Gaussian mixture model as model of individual differences to better understand the patterns in how participants vary in terms of their exploratory strategies. Four families of strategies emerged from the data across all four experiments. We analysed the behaviour of participants across the four

experiments presented in Chapter 2 according to these four clusters of participants and found consistent behavioural characteristics. This showed that there existed systematic similarities and differences between participants across the different experiments. In general, we find that the behavioural qualities of each sub-group were coherent with their parameter descriptions. We named the largest cluster of participants *Maximisers*. Maximisers had the best overall performance, and progressed across grids. They were predominantly explained by selecting utility driven actions, relying on generalisation to predict the rewards of unseen actions. This was characterised by small exploratory steps in the direction of the maximum and a re-selection of the maximum once found. The second best performing group was the *Scholar* sub-group. It was the only cluster with a significant weight on *uncertainty directed search*, which was characterised by larger average distances between initial selections of participants. Like the Maximiser group, they also displayed progress across grids. The remaining two groups were predominantly explained by the heuristic components of the model. In general, *Greedy local* participants improved their scores across turns, but there was no evidence for transfer across grids. The most important term defining this group of participants was the greedy-re-selection component, which was characterised by participants settling on sub-optimal actions. Finally, the last cluster of participant strategies we identified was the one of *Local explorer* participants. Local explorers explored almost exclusively, and did so by favouring local and novel actions. Though they often found the maximum, they did not show strong progress across trials, nor did they progress across grids. Overall, these results paint a picture of human exploration as diverse, yet with systematic mechanisms underlying it. The majority of participants were in parts explained by the expected utility of actions, showing that people rely on generalisation to predict the outcome of unknown actions. Only a select number of participants were explained by uncertainty driven search. Many participants seemed to rely on local search to guide their exploration, though his may be have been adaptive to the nature of our task (i.e.

spatially-correlated, and uni-modal with a steep gradient). In the next chapter, we look at experiments conducted by Wu *et al.* (2018) to understand if our findings are coherent with their data. Their experiments carried many similarities with the experiments we presented in Chapter 2, but with different reward structures.

Chapter 5

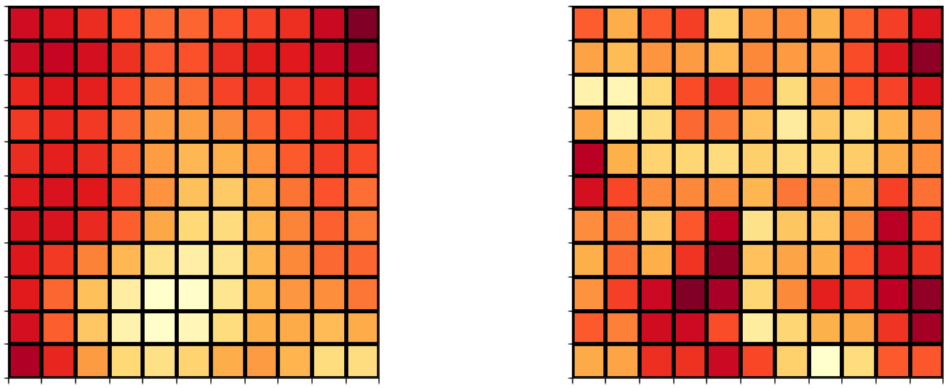
Studying generalisation in rough and smooth environments

5.1 Introduction

We have reported, through our general model of human exploration, consistent patterns in the different strategies used by participants across four different experiments. One of the limitations from our study is that in all four experiments, participants were presented with a single type of reward structure. The grids had one maximum, and the rewards of tiles would decay exponentially with the Euclidian distance to the maximum tile. In this chapter, we look at participant data collected by Wu *et al.* (2018)[1]. Their experiments carried many similarities with the experiments we presented in Chapter 2, but used different reward structures. Here, we focus specifically on the influence of environment structure on the strategies used by participants. Our goal in this chapter is two-fold. First,

[1]Participant data, code and model simulation data from their study are available at https://github.com/charleywu/gridsearch.

we look at individual differences, and the patterns of variation in strategies used across participants. Can the patterns highlighted in our analysis in Chapter 4 be observed in how people search in more complex (multi-modal, and less spatially correlated) sets of reward structures? Part of this investigation lies in understanding to what extent people's strategies are adaptive to the underlying structure of their environment. Second, we examine the different claims made by Wu *et al.* (2018) and relate them to our findings from the previous chapters. We start by reviewing briefly the experiments presented in their study and how they differ from the experiments we presented in Chapter 2.



(a) Smooth reward structure **(b)** Rough reward structure

Figure 5.1: Examples of grids with smooth and rough reward structures presented to participants.

5.1.1 Summary of experiments

In the study conducted by Wu *et al.* (2018), participants were presented with eight grids of 11x11 tiles. Wu et al. contrasted two conditions, one where participants had to search for the maximum and one where they had to maximise their cumulative score. Here, we look at the experimental conditions where participants

had to maximise their cumulative reward across all grids, a goal which was identical to our tasks'. We also restrict our analysis to their second experiment (Experiment 2), where the generating parameters of the reward structures are known. Wu *et al.* (2018) manipulated the types of reward structures with *smooth* versus *rough* distributions of rewards, corresponding to the correlations between the location of a tile and its associated reward (see Figure 5.2). The reward values were sampled from RBF kernels with different length-scales (λ=1 for rough and λ=2 for smooth). In their formulation, they set the lengthscale as $\lambda = \frac{1}{2}l^2$, the denominator of the RBF kernel function for a more psychologically interpretable formulation. The prior mean was fixed to the median value of payoffs, $m(\mathbf{x}) = 50$.

In the smooth condition, high rewards were strongly spatially correlated, indicating that tiles close to each other had a high degree of similarity in their associated rewards, while a rough environment presented participants with more unpredictable outcomes (i.e. a lower degree of correlation between neighbouring tiles). Across eight grids, each participant was presented with alternating search horizons (20 trials vs 40 trials) in both experimental conditions with the order counterbalanced between subjects. For simplicity, we focus here on the long horizons, i.e. four of the grids presented to participants. One reason for this is to look at whether people's tendency to over-explore, as observed in our experiments, was not set up by the relatively short search horizons in our experiments (20 selections). While we do not report on the analysis for shorter horizons here, we found our broad conclusions to apply across the two different types of tasks.

Beyond the reward structures, there were two main differences from our experimental design. First, participants were given a cue about the maximum, as the rewards were colour hued, with the maximal value being dark red. Knowing the the maximum of the function offers a considerable advantage to the search – this can be seen through the efficiency of Bayesian Optimisation methods that seek to directly minimise the entropy of the maximum y-value (e.g., see Wang & Jegelka,

2017). Additional information about the maximum value may have motivated more exploitative selection strategies in participants. Second, there was some added uncertainty with the rewards of individual tiles, as a small noise value was added upon selection ($\epsilon \sim \mathcal{N}(0, 1)$, with 100 being the maximum value). The rewards were scaled in each grid so the displayed reward values would be different. The history of previously observed rewards remained accessible throughout every given grid.

Across both experimental conditions, participants were initially shown three examples of grids with the same correlation between rewards as in the real task (i.e. sampled from the same RBF kernel). Participants were also given written instructions, comprehension check questions, and feedback between rounds to ensure they were familiar with the underlying reward structures (see Wu et al.'s Methods section for details).

5.1.2 Results from Wu *et al.* (2018)

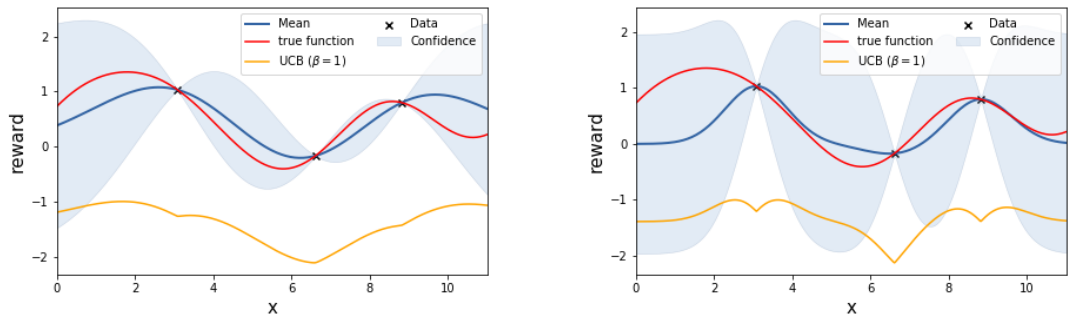
In their study, the authors draw some conclusions about human behaviour that we either did not investigate directly, or are at odds with the ones we reported in the previous chapters. We review them briefly here:

1) People have a systematic tendency to under-generalize the extent of spatial correlations. Wu *et al.* (2018)

When predicting the expected reward of unseen actions, participants need to evaluate the similarity between a new action and previous observations. The general assumption is that actions close to each other will be similar, while distant ones will be more unpredictable. In the context of the grid tasks, the degree to which participants generalise will inform how similar one expects neighbouring

tiles to be from an observation when they are e.g. two, three or four tiles away. In a GP model, the length-scale parameter can be interpreted as determining a "generalisation gradient" (cf. Shepard, 1987), as it models generalisation as a (squared) exponential decay distance between stimuli. The larger the length-scale, the smoother the reward function – thus predicting nearby actions to be more similar.

In their study, Wu *et al.* (2018) put forward the idea that people tend to assume a lesser degree of similarity between actions than the true degree of similarity of the environment. Assuming a lower degree of similarity predicts that people are more uncertain about unobserved actions than an ideal observer. To give an intuition for this, we show an example using a Gaussian Process model for two different length-scale values ($\lambda = 2$ and $\lambda = 0.8$). We look at the different model predictions after three observations from a function sampled from a GP kernel with length-scale 2.



(a) GP predictions for $\lambda = 2$.

(b) GP predictions for $\lambda = 0.8$

Figure 5.2: GP predictions for different length-scales (λ). The yellow line shows the UCB predictions based on the GP predictions for β=1. The UCB acquisition function favours neighbours to previous observations when the length-scale is smaller.

We can see from the figures that a smaller length-scale leads to greater overall uncertainty about the underlying function. This influences a UCB acquisition

function towards favouring both *new* actions, and actions that are *close* to known and highly rewarding actions. These predictions seem to broadly align with the local search behaviour observed in our experiments.

Our modelling approach did not attempt to fit the length-scale to participants. Instead, we presented a GP learning model which estimated the posterior over kernel parameters after each new observation. The uncertainty about kernel hyperparameters was then integrated out when computing the acquisition function. In our initial two experiments participants were not familiar with the underlying structure and had to learn it through their observations. In this context, the idea of a learning model implies that the hidden reward structure is unknown, and is thus at odds with a fixed length-scale across all trials. A fixed length-scale does however make a lot of sense in the context of known structures. By assuming that the true parameter has already been learned during training, it can be interpreted as a prior over the expected reward correlation. This is the interpretation we opt for in this chapter. We investigate in this chapter whether the local bias of participants can be explained by a greater uncertainty assumed by participants.

2) There is substantial evidence for the separate phenomena of directed exploration (towards reducing uncertainty) and noisy, undirected exploration. Wu *et al.* (2018)

3) Human search behaviour is best explained by Gaussian Process function learning combined with an optimistic upper confidence bound sampling strategy. Wu *et al.* (2018)

The picture painted through claim 2) is that people explore through a combination of uncertainty reducing actions and random actions. 3) emphasises the important role of uncertainty in the decisions of participants. When taken at face

value, this is largely at odds with what we found in our experimental data. Across our experiments, only a small proportion of participants relied on structural uncertainty to guide their search (the group we named scholars). Furthermore, there was no strong evidence for random search, but rather local search with a strong bias towards novel actions. As explained above, assuming a tendency to under-generalise, the theory put forward by Wu *et al.* (2018) may partly explain what we observed in our data, specifically local search strategies and a strong exploratory drive. In this chapter, we attempt to clarify their psychological claims by looking at the predictions made by their model. We first highlight some points that lack clarity in the results reported by Wu *et al.* (2018).

In their study, Wu *et al.* (2018) contrast different versions of the same models, a standard model and a "localised" version. The localised version of a model consists in multiplying the predictions of a model by their Inverse Manhattan Distance (IMD) to the previous selection. While they report that Gaussian Process regression combined with UCB sampling provides the best account for how people explore, 61 participants out of 82 were best predicted by the localised version of their model. The claim that participants rely on generalisation to guide their search is based upon a comparison between a function learning model (that uses a GP for generalisation) and an option learning model, which simply learns the value of previously observed tiles without extrapolating to unseen ones. However, the localised version of the option learning model strongly outperforms the non-localised GP-UCB model ($t(79) = 10.47, p < .001$). This indicates that the individual contributions of the localisation component and of the model based components (i.e. the expected reward and the uncertainty under the GP) need to be investigated in more detail. In this chapter, we reanalyse the predictions of the different Gaussian Process models, and compare it to our general model to better understand the extent to which the local bias influences the inferred

parameter of the model when fit to participants. Finally, when comparing the length-scales Wu *et al.* (2018) obtained by cross-validation for participants on long horizons, we find no significant difference between the rough and smooth conditions (Δ_M=0.08, $t(79) = 0.95, p = 0.35$). This would indicate participants assumed the same degree of similarity in the two different conditions and did not adapt their representations according to the structure of the environment. A possible interpretation is that participants generalise in a very localised way, as opposed to using the global structure of their environment to guide their search efficiently. To further investigate people's ability to learn and exploit the task structure they are faced with, we look at the inferred generalisation parameters through the lens of our general model.

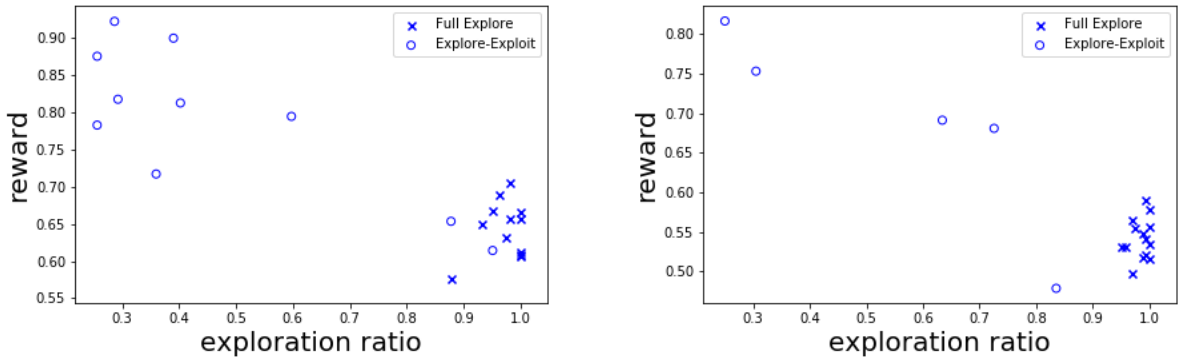
Before looking at the predictions of different models, we first present a re-analysis of their experimental data and look at patterns of individual differences.

5.2 Individual differences: an empirical analysis

One of the salient observations from our experimental analysis in Chapter 2, was the significant number of participants who exclusively explored despite the clear incentives to repeatedly select the most rewarding tiles. This pattern of behaviour was not reported in Wu *et al.* (2018), but some of the learning curves of individual participants (shown in their supplementary materials) hinted that this may be the case. We look at whether the pattern of purely exploratory behavior that we found is also present in Wu et al.'s data. To assess the exploratory drive of participants, we analyse the ratio of unique selections, and how it relates to participant performance. We separate participants into *Full-explore* and *Explore-exploit* groups according to their ratio of exploration like we did in Chapter 2, where exploration is defined as the selection of a previously unobserved action.

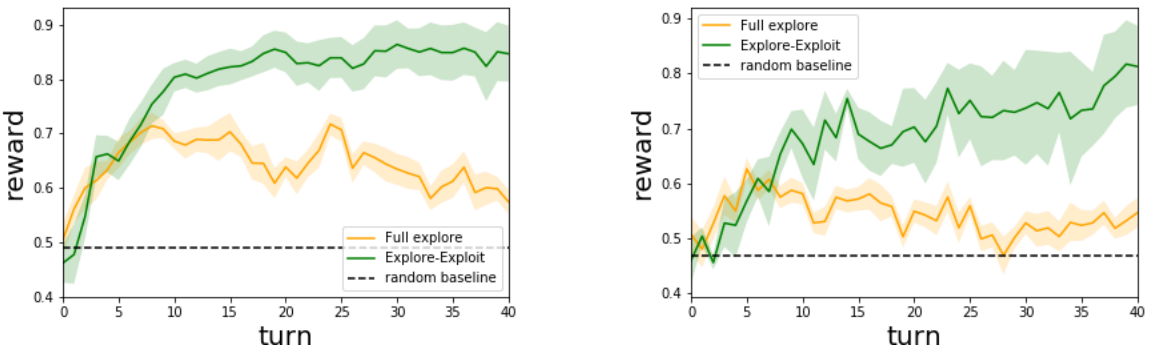
We define a *Full-explore* strategy as selecting unique tiles more than 0.90 of the time (at least 37 out of 40 selections) on the majority of grids (at least 3 out of 4). We chose these values to be both conservative about what was considered fully exploratory behaviour and similar to what we used in our own experiments. Figure 5.3 shows participants' average performance in relation to their explore-exploit ratio. In the Smooth condition, 0.55 of participants had a *Full-explore* strategy (12 out of 22). In the rough condition, it was 0.74 of participants (14 out of 19). This was not a significant difference under Fisher's exact test ($p = 0.33$). In Figure 5.4, we plot the learning curves of participants according to this grouping of participants. As reported by Wu *et al.* (2018), participants in the smooth condition performed significantly better than in the rough condition ($t(42) = -4.26, p < 0.001$). This supports the assertion that participants were able to exploit the underlying reward structure, since the smooth condition was more regular. However, it does not necessarily imply participants *learned* the structure, since a simple gradient-ascending line-search would do better in a smoother space. From our analysis, it is also clear that there were important differences in participant strategies. More specifically, we observed a similar clear partition in the amount of exploration conducted by individual participants than the one we described in Chapter 2. A large proportion of participants explored exclusively, visibly dismissing rewards, while the others traded off between exploration and exploitation so as to maximise rewards.

In accordance with the local bias in participant selections observed in our experiments, Wu *et al.* (2018) reported that participants sampled more locally than a random baseline in both conditions. The average distance between exploratory selections (using the Inverse Manhattan Distance, IMD) was smaller in the smooth condition (M=2.08, SD=0.40) than in the rough condition (M=2.39, SD=0.70), though not significantly different (t(42)=1.68, p=0.10). One could have expected participants in the smooth condition to have more global selections since the



(a) Smooth reward structure **(b)** Rough reward structure

Figure 5.3: Average reward performance with respect to the explore-exploit ratio of participants in the Rough and Smooth conditions. A value of 1 indicates exclusively selecting new tiles. Each dot represent a single participant.



(a) Smooth reward structure **(b)** Rough reward structure

Figure 5.4: Participant performances according to Full-Explore and Explore-Exploit strategy types in the Smooth and Rough experimental conditions.

structure was simpler to exploit, but it could be that participants opted instead for small steps, ascending the gradient, much like what we observed in our experiments. Participants could have taken more random exploratory steps in the rough condition since the structure carried more inherent uncertainty.

In summary, we analysed the participant presented in Wu *et al.* (2018) to better understand the influence of the environment structure on participant strategies. Participants were able to perform better under the smooth condition, showing that they were able to use the reward structure to maximise their rewards. Although this was not a point of focus in the study conducted by Wu et al., we found important individual differences amongst participant strategies. Specifically, some of the patterns reported in Chapter 2 were also salient in the behaviour of participants. Indeed, a large proportion of participants engaged in full exploratory strategies in both the rough and smooth conditions, failing to re-select known rewarding tiles. Furthermore, participants also displayed a strong local bias in their exploratory selections. In general, these observations seem in conflict with some of the descriptions made by Wu et al., when e.g. noting a "remarkable concurrence between intuitive human strategies and state-of-the-art machine learning research", or the importance of uncertainty estimates to guide exploration. To better understand the different strategies used by participants, we use the general model introduced in Chapter 3 in the next section.

5.3 Model based analysis of participant strategies

We report here on the group level parameter distributions obtained from the general model. In our initial model based analysis, we fix the length-scale to the true generating parameter of the reward structure instead of estimating the

posterior over GP parameters. This is done to evaluate how participants might rely on the true structural uncertainty, given correct assumptions. We use a fixed parameter value under the assumption that participants did not have to learn the correlation of rewards, since they had seen examples of the reward structure, and were explicitly told about it. We scale the predictive scores of each component by their range across all grids and trials to get a better understanding of their respective contribution when explaining participant behaviour (for details on the model, and model fitting, see Section 3.3.

5.3.1 Participant strategies in smoothly spatially-correlated environments

In the smooth condition, no participants had a contributing β parameter (M=0, SD=0), indicating that participants did not rely on global structural uncertainty to direct their search (see Figure 5.5). All participants were partly explained by the expected rewards under the GP (M=0.57, SD=0.32), and 8 participants were almost exclusively explained by it (M=0.99, SD=0.02). The other 14 participants had an important local search term (M=0.41, SD=0.06) and novelty term (M=0.22, SD=0.1). None of the participants had a significant greedy component (M=0.03, SD=0.05). In general, these results seem coherent with two of the strategy groups identified in Chapter 4: *Maximisers* and *Local explorers*. One possible explanation for the absence of *Scholars* was that participants were trained on the reward structure prior to the tasks so opted for local gradient ascent steps. It could be that the nature of the reward structure made it easier for participants to predict expected rewards and thus led more participants to a *Maximiser* strategy as opposed to a *Greedy local* one.

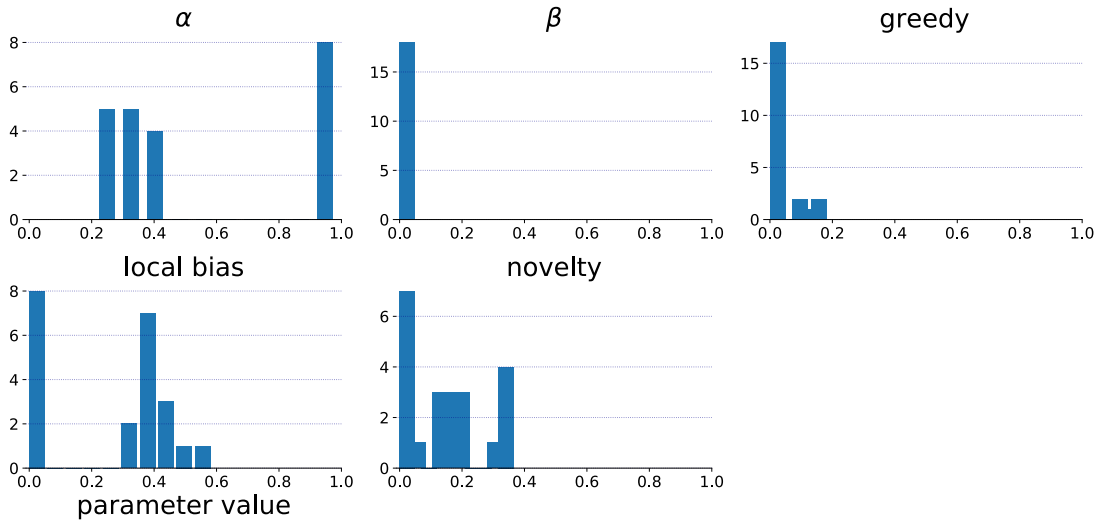


Figure 5.5: Histograms over parameter values for participants in the smooth condition.

5.3.2 Participant strategies in the rough condition

In the rough condition (n=19), three participants had a non-zero β parameter, though this was practically insignificant (M=0.05, SD=0.01) (see Figure 5.13). Similarly, four participants had non-zero greedy parameters, but these contributed only modestly to explanations of their behaviour (M=0.01, SD=0.02). All participants were explained by reward driven actions, as modelled by the expected mean under the GP in the general model (M=0.74, SD=0.14), more so than in the smooth condition ($U(42) = 142.0, p = 0.07$). Local search (M=0.15, SD=0.11) and the novelty component (M=0.10, SD=0.05) were also important contributors for participants. The local component contributed on average significantly less than in the smooth condition ($U(42) = 138.0, p = 0.05$). The novelty term was also less important in the rough condition than in the smooth, but not significantly so ($\Delta_{Med} = 0.05, U(42) = 162.5, p = 0.17$). In general, the strategies in the rough condition as described by the general model seem to correspond to the ones observed in the smooth condition, though with a higher tendency for reward

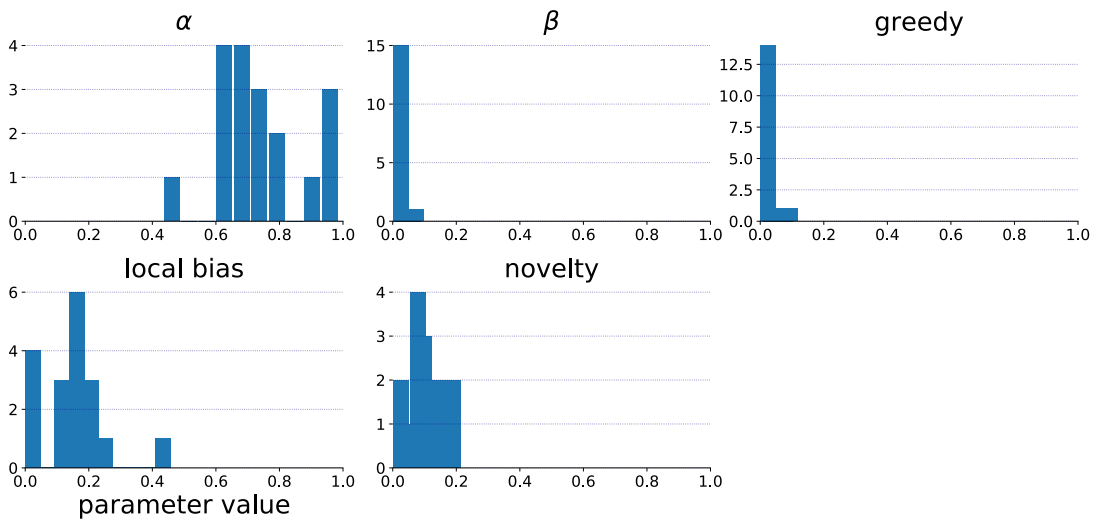


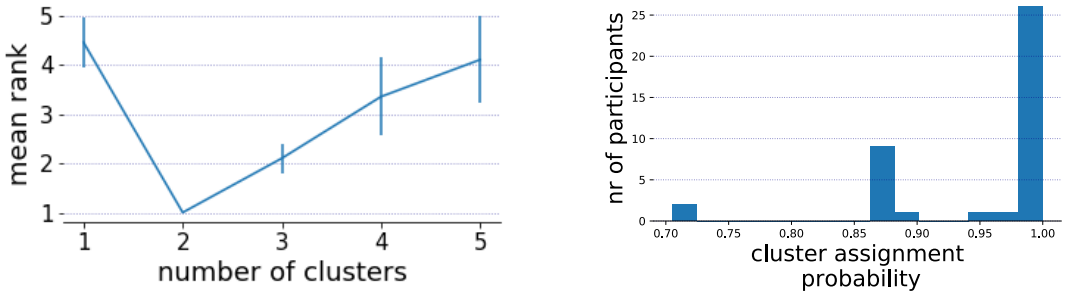
Figure 5.6: Histograms over parameter values for participants in the rough condition.

driven actions, and less local search. We look at the differences in strategies used by participants in the next section.

5.3.3 Identifying clusters of participant strategies

To identify patterns in the different strategies used by participants, we use the same individual differences model that we did in Chapter 4, including a Gaussian Mixture Model to identify clusters in the MLE parameters fit to participant selections. As in Chapter 4, we use leave one out cross-validation to estimate the number of clusters

The mean negative log likelihood was estimated for a range of groups K=1,. . . , 5 in each cross-validation fold. We removed one participant whose weight parameters were zero from the participant data set under the assumption that random/unpredictable behaviour was an independent cluster of behaviour. On average, the cross-validation of the GMM indicated that a separation into two



(a) Average rank of GMM with different number of clusters when comparing scores on test set using 6-fold cross validation. The results shown are from 20 evaluations.

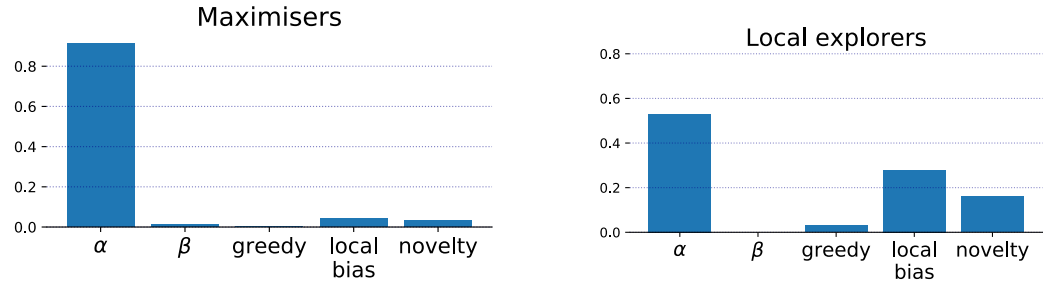
(b) Cluster assignment probabilities across participants for K=2 clusters.

Figure 5.7: GMM cross-validation results and cluster assignment probabilities with K=2 clusters.

sub-groups provided the most generalisable model, corresponding to the lowest negative log-likelihood score on the test set (see Figure 5.7).

When looking at the cluster assignment probabilities, all participants were assigned to one of the two clusters with an average probability of 0.95 (SD=0.1). The cohesion of clusters supports the hypothesis that distinct families of strategies amongst participants exist. We show in Figure 5.9, 5.10, 5.11 and 5.12 examples of participants belonging to the two sub-groups in each condition. We highlight in particular the similarity with the strategies observed in Chapter 4, both in the cluster parameters and in the qualitative patterns of behaviour exhibited by participants. Accordingly, we refer to the participants belonging to two cluster of participant strategy as Maximisers and Local explorers.

In the smooth condition, there were 15 participants clustered under the Local explorer group, and 7 under the Maximiser group. In the rough condition, there were 13 local explorers and 5 Maximisers. Bar 4 exceptions across both



(a) (b)

Figure 5.8: Cluster centre parameters obtained from the GMM clustering algorithm.

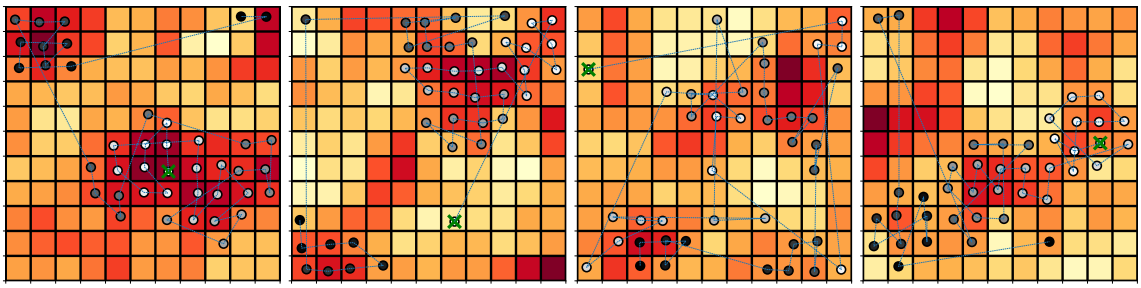


Figure 5.9: Example of of a *Local explorer* participant in the rough condition.

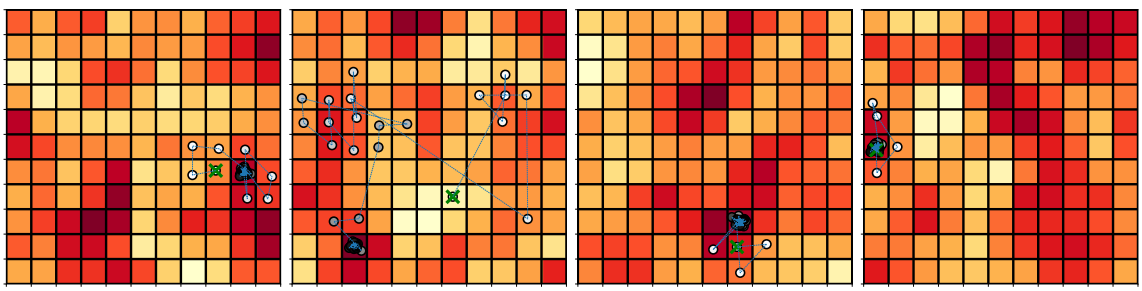


Figure 5.10: Example of a *Maximiser* participant in the rough condition.

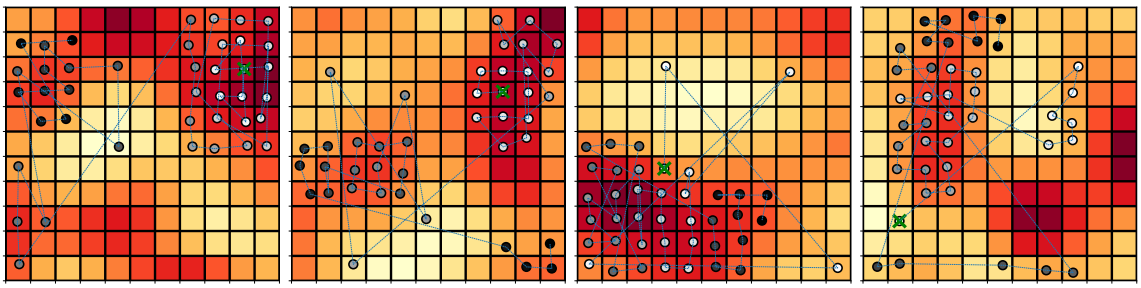


Figure 5.11: Participant clustered under the Local explorer group in the smooth condition.

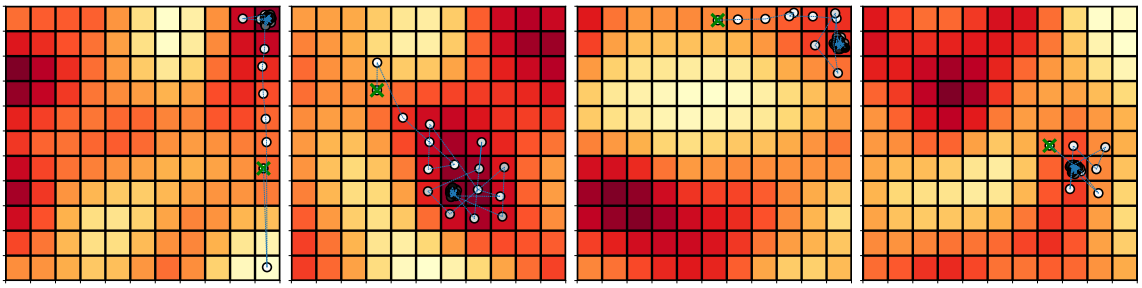


Figure 5.12: Participant clustered under the Maximiser group in the smooth condition.

conditions, this corresponded roughly to the Full-explore and Explore-exploit dichotomy presented in Section 5.2.

5.3.4 Preliminary discussion

So far, we have analysed the strategies used by participants across two experimental conditions: one where the correlation between location and reward is rough, and one where it is smooth. By conducting an empirical analysis of participant behaviour, we found that within both conditions there were significant differences amongst individuals. Like in our experiments, we found that a significant proportion of participants exclusively selected exploratory actions. This was the case despite the fact that participants were familiar with the spatial correlation of rewards, knew the relative value of their actions (i.e. how close to the maximum it was), and the high number of trials they were given (40 selections per grid). To better understand participant strategies, we used the general model presented in Chapter 3. From our model based analysis, two group of strategies emerged. These strategies corresponded neatly to two of the ones described in Chapter 4, in our own experiments: *Local explorers* and *Maximisers*. Local explorers, who engaged in highly exploratory behaviour, principally relied on local search, were influenced by the expected value of actions, and were driven by novelty. Maximisers traded off between exploration and exploitation, meaning they converged on a high value after a phase of exploration and re-selected it, and were predominantly guided by the expected value of actions. Maximisers significantly outperformed participants in the Local explorers group. Across both condition, there was no evidence for participants relying on uncertainty to guide their search.

At the beginning of this chapter, we presented two psychological claims made by Wu *et al.* (2018). The first claim is that participants have a tendency to under-generalise. This was not investigated directly so far, though there

was strong evidence for local bias in the way participants explored. As we discussed in the introduction, it is possible that local exploration may be partly influenced by having a prior favouring lower degrees of correlation between action rewards. The second claim is that participants rely on both directed exploration (towards reducing uncertainty) and random exploration. This was not supported by our model-based analysis, as no participants were explained by our model's uncertainty driven exploration component. Instead, local search, actions driven by their expected rewards, and a bonus for novel actions were the three most important components to describe how participants selected their actions. In section 5.5, we compare the predictions of our model to the one presented by Wu et al. and show that it offers better predictive power over participant selections.

The notion that people rely on generalisation to guide their search means they construct a representation of the environment, where the outcome of unseen actions is informed by previous observations, and use this model to inform their actions. In the context of a rational analysis, when trying to understand the computational problem people are solving, it suggests the way participants select actions is adaptive to the structure of their environment. For example, if guided by uncertainty, a smoothly correlated environment would encourage participants to explore more globally, while rough correlations would encourage more local exploration due to the higher levels of uncertainty. Our model-based analysis did not investigate to what extent participant were able to learn an accurate representation of the environment, but simply assumed participants had correct expectations about the correlation structure of the grids. One difference we did find between the two experimental conditions was that participants in the smooth condition explored more locally than participants in the rough condition. In the next section, we relax the assumption that participants knew the correct correlation structure and investigate participants' ability to adapt to the structure of their environment during search.

5.4 Generalisation in search

To understand how adaptive people's strategies are to their environment, we first look at the predictions of the general model when having the length-scale as a free parameter. Fitting the length-scale can be interpreted as capturing participants' expectation about the correlation between rewards in the grid, or to what extent they generalise between nearby actions. So far, our models have relied on an ideal observer analysis that evaluates to what extent people reduce global uncertainty given that people hold correct assumptions about the underlying reward structure. Our model results did not support the hypothesis that people aim to reduce global uncertainty during search. Instead, people relied on local search and the expected value of rewards. We now test the hypothesis that people have a tendency to under-generalize, and that their expectation of rougher reward structures leads them to select actions aimed at reducing local uncertainty.

We use the AIC to compare the model fits of the general model when fitting the length-scale to participants against the general model where it is fixed. Fitting the length-scale gave better AIC scores for all participants, both in the rough and the smooth condition. The difference in scores was more pronounced in the smooth condition (Mdn=51.22) than in the rough condition (Mdn=12.74) ($U(42) = 97.0, p = 0.002$). To understand the contributions of the individual components, we scale the parameters according to their range and normalise them. Again, we find that uncertainty directed search was not a principal component to explain participant behaviour. Only three participants had a non-zero β parameter (M=0.14, SD=0.06) in the rough condition, and two in the smooth condition ([0.18, 0.11]).

Next, we look at the length-scale for all participants. The length-scale was a meaningful parameters across participants since they were all explained by

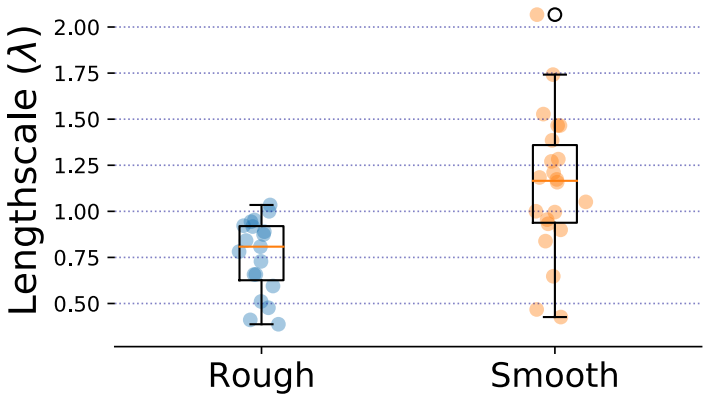


Figure 5.13: Distribution over length-scale parameter values for participants in the rough ($\lambda = 1$) and smooth ($\lambda = 2$) condition. Participants were able to adapt to the structure of the environment and used larger "generalisation gradients" in the smooth condition.

the expected reward under the GP with a value of at least 0.16. This is compatible with the results found earlier, indicating that participants relied on generalisation to guide their search. We find that in both conditions participants had smaller length-scales than the true generating parameter value. In the smooth environment ($\lambda = 2$), the average length-scale fit to participants was of 1.14 (SD=0.38). In the rough environment ($\lambda = 1$), the average length-scale was of 0.76 (SD=0.20). The length-scale fit to participants was significantly smaller in the rough environment than in the smooth environment ($t(42) = -3.89, p < 0.001$).

In general, these results support the hypothesis that participants had used generalisation to guide their search, had a tendency to under-generalise, and the extent to which they generalised was adaptive to the structure of their environment. However, there was little evidence to support that participants relied on uncertainty to guide their search, even when accounting for participants' tendency to under-generalise.

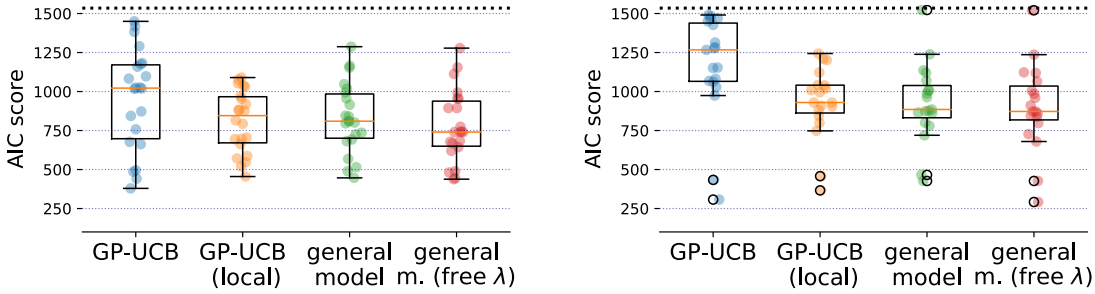
To better understand the psychological claims made by Wu *et al.* (2018) in relationship to the predictions made by their model, we compare our general model with λ as free parameter against their GP-UCB model, a well as its localised variant (GP-UCB*).

5.5 Model comparison

It may be important to point out that the GP-UCB model was nested within the general model (with free length-scale parameter), whereas its localised variant (GP-UCB*) was not. The local component of the GP-UCB* model was multiplied to the GP-UCB predictions, making it more difficult to tease their respective contributions apart, while the local component of the general model was added to the predictions of the other components.

We compare the AIC scores of the general model to the GP-UCB model, and its localised variant presented by Wu *et al.* (2018). The general model gave significantly better AIC scores than the GP-UCB model ($\Delta_{Med} = -248.25, U(83) = 465.0, p < 0.001$). Similarly, GP-UCB* produced significantly better AIC scores than the vanilla GP-UCB model ($\Delta_{Med} = -178.6, U(83) = 496.0, p < 0.001$). The general model gave better scores than GP-UCB* (Δ_{Med}=-69.63), though this was not significant ($U(83) = 734.0, p = 0.16$). Overall, 26 participants were best predicted by the general model (with free λ), against 12 by the GP-UCB* model, and 3 by the Vanilla GP-UCB model (see Figure 5.14).

To further evaluate the models, we look at the out of sample predictions of the models by using the leave-one-out cross-validation procedure. Leave-one-out cross-validation offers the benefit of directly estimating the predictive accuracy of the model without having to arbitrarily penalise model complexity, contrary to methods such as the AIC or BIC. This is done by iteratively fitting the



(a) AIC scores for participants in the smooth condition.

(b) AIC scores for participants in the rough condition.

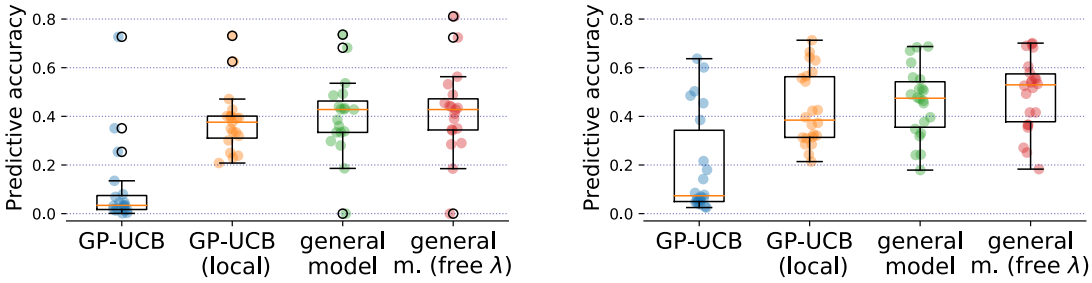
Figure 5.14: Boxplots show the median (line), interquartile range (box),and 1.5x IQR (whiskers). Each individual participant is represented as a dot. The dotted line shows the random baseline.

models on three grids, and generating out of sample predictions on the remaining grid. We compare the prediction error by summing the log loss over all rounds. Following Wu *et al.* (2018), we use a pseudo-R^2 measure that compares the total log loss prediction error for each model to that of a random model:

$$R^2 = 1 - \frac{\log \mathcal{L}\left(\mathcal{M}_k\right)}{\log \mathcal{L}\left(\mathcal{M}_{\text{rand}}\right)}$$

where $\log \mathcal{L}\left(\mathcal{M}_{\text{rand}}\right)$ is the log loss of a random model (i.e., picking options with equal probability) and $\log \mathcal{L}\left(\mathcal{M}_k\right)$ is the log loss of model k's out-of-sample prediction error. Intuitively,$R^2 = 0$ corresponds to prediction accuracy equivalent to chance, while $R^2 = 1$ corresponds to theoretical perfect prediction accuracy, since $\log \mathcal{L}\left(\mathcal{M}_k\right) / \log \mathcal{L}\left(\mathcal{M}_{\text{rand}}\right) \to 0$ when $\log \mathcal{L}\left(\mathcal{M}_k\right) \ll \log \mathcal{L}\left(\mathcal{M}_{\text{rand}}\right)$. R^2 can also be below zero when the model predictions are worse than random chance.

Here, the general model generated the best predictions for 27 participants, while 13 participants were best predicted by the GP-UCB* and one by the general model with fixed λ (see Figure 5.15). The general model provided significantly better predictions across participants ($\Delta_{Med} = -0.067, U(83) = 642.5, p = 0.03$).



(a) R scores for participants in the rough condition.

(b) R scores for participants in the smooth condition.

Figure 5.15: Boxplots show the median (line), interquartile range (box),and 1.5x IQR (whiskers). Each individual participant is represented as a dot. The dotted line shows the random baseline. Higher score is better.

In addition to the better performance of the general model, this result shows that model comparison measures that take into model complexity (such as the AIC above) may not be the most appropriate, especially when using regularisation methods in the optimisation like we did for the general model. Indeed, a large number of participant AIC scores had a complexity penalty for non contributing parameters that were driven to zero due to the L1 penalty, while leave-one-out cross validation remains agnostic to the number of parameters.

In summary, we have found that our general model is robust in predicting participant performance, and was the best predicting model for most participants. The analysis of the general model parameters, when fit to participants, showed that participants did not rely on uncertainty to guide their search. Instead, they mainly relied on the expected reward of actions, and local search paired with a novelty drive. This is further supported by the fact that the vanilla GP-UCB model was significantly worse at predicting participant behaviour than the other models considered. Comparatively, the localised version of the UCB-model (GP-UCB*) offered a significantly better predictive account than the non-localised one, and relied on the uncertainty parameter β to make robust predictions. In the next

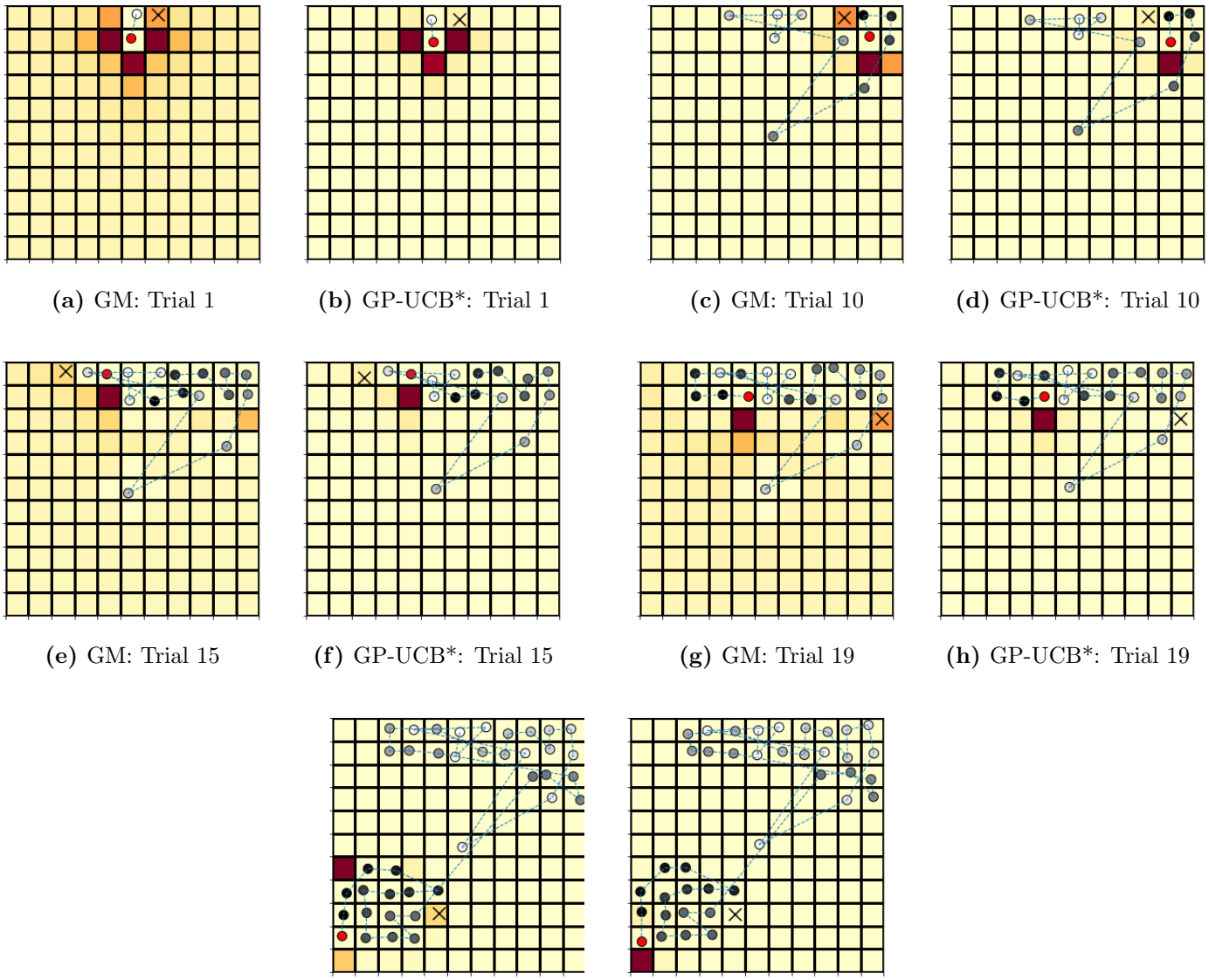
part we inspect visually the predictions of the localised GP-UCB model to better understand its semantics and the psychological claims it carries. We select trials where the predictions of the general model and the UCB-model* were distinct to understand their differences.

5.6 Investigating model predictions: Local uncertainty as heuristic

In this section, we compare the predictions made by our general model, and the GP-UCB* model advocated by Wu *et al.* (2018). The first participant was better predicted by the general model (see Figure 5.17), while the second was better predicted by the GP-UCB* model.

The first participant was fit with important contributions from the expected reward term (α=0.39), local bias term (λ=0.29) and novelty component (ν=0.31). The MLE under the GP-UCB* model gave a softmax value of τ=0.10, an length-scale parameter of 0.62 and a β parameter of 0.58.

We find that the predictions made by the GP-UCB* model are exclusively direct neighbours to the previous selection. In the case of the first participant, this prevents the model from capturing some of the exploratory actions, where the participant goes back to a distant tile near a previous selection (e.g. Trial 10, 15, 19, 38). In this case, the predictions of the general model are more flexible, and seem to have a more nuanced model of uncertainty. In the case of the second participant, the fact that the GP-UCB* puts weight on fewer actions leads to better predictions overall (e.g. Trial 10, 15, 19), though this leads to few poor predictions of participant selections (e.g. Trial 1 and 38).



(a) GM: Trial 1 **(b)** GP-UCB*: Trial 1 **(c)** GM: Trial 10 **(d)** GP-UCB*: Trial 10

(e) GM: Trial 15 **(f)** GP-UCB*: Trial 15 **(g)** GM: Trial 19 **(h)** GP-UCB*: Trial 19

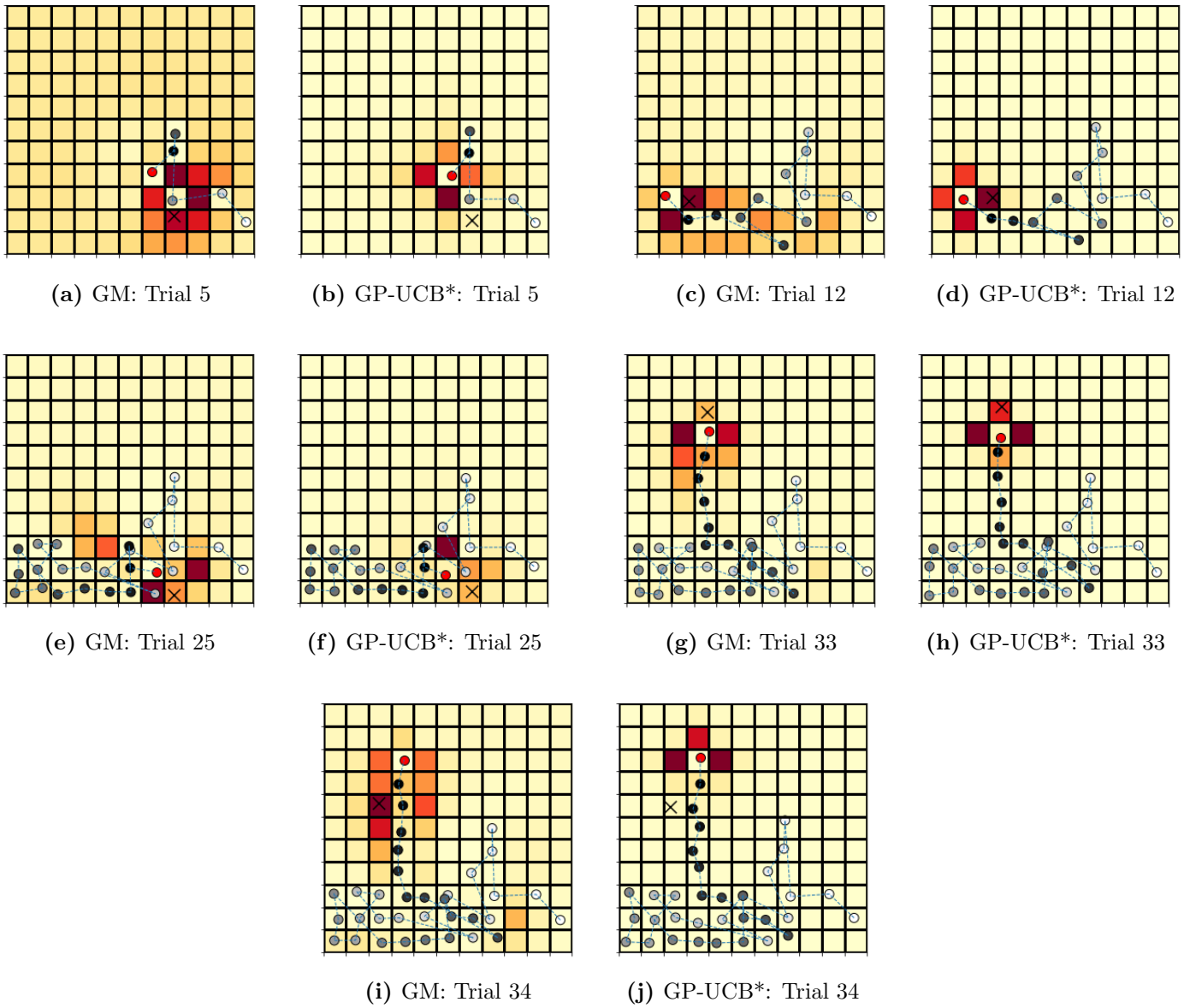
(i) GM: Trial 38 **(j)** GP-UCB*: Trial 38

Figure 5.16: Model predictions of a participant in the smooth condition at trial 1, 10, 15, 19 an 38. GM indicate predictions made by the general model. GP-UCB* are predictions made by the localised GP-UCB model. The hue of a tile indicates the probabilities predicted by the model. The red dot indicates the last selection, the cross indicates the upcoming selection (the one to be predicted by the model). The hued dots indicate previous participant selections, with a darker hue indicating more recent selections.

The visual inspection of the predictions by the GP-UCB* model make it clear that it is a model of local search, informed by local uncertainty and expected rewards. Its predictions are very much like the line-search heuristic we discussed in Chapter 2 and 3. The overall good predictions of the model when fit to participant selections could suggest that participants rely on *local* uncertainty to guide their search. This tendency to favour local uncertainty over global uncertainty has been reported in active learning setting where people had to learn category boundaries (Markant *et al.*, 2016b), and could have several benefits in terms of memory constraints and computational resources. One explanation could be that local uncertainty is informative, and cheaper to compute than global uncertainty. A similar phenomenon has also been found in causal learning, showing that participants prefer local updates of their beliefs while keeping the same general hypothesis instead of more sudden "global" updates (Bramley *et al.*, 2017). This preference for locally informative actions over actions aiming to reduce "global" uncertainty has also been reported in the domain of visual search tasks (Renninger *et al.*, 2007).

5.7 Conclusion

In this chapter, we set out to investigate the importance of the environment structure for participants' search strategies. To do this we analysed data from two experimental conditions collected by Wu *et al.* (2018). In one environment, the rewards were smoothly correlated, while in the other the correlation between the outcome of similar actions was rough. In general, Wu *et al.* (2018) noted a "remarkable concurrence between intuitive human strategies and state-of-the-art machine learning research", with "the vast majority of participants best described by the Function Learning-UCB model or its localized variant". Instead, we found that the behaviour of participants carried many similarities to the behaviour



(a) GM: Trial 5 **(b)** GP-UCB*: Trial 5 **(c)** GM: Trial 12 **(d)** GP-UCB*: Trial 12

(e) GM: Trial 25 **(f)** GP-UCB*: Trial 25 **(g)** GM: Trial 33 **(h)** GP-UCB*: Trial 33

(i) GM: Trial 34 **(j)** GP-UCB*: Trial 34

Figure 5.17: Model predictions of a participant in the smooth condition at trial 5, 12, 25, 33 an 34. GM indicate predictions made by the general model. GP-UCB* are predictions made by the localised GP-UCB model. The hue of a tile indicates the probabilities predicted by the model. The red dot indicates the last selection, the cross indicates the upcoming selection (the one to be predicted by the model). The hued dots indicate previous participant selections, with a darker hue indicating more recent selections.

observed in our experimental data from Chapter 2. Specifically, we found similarly salient patterns of individual differences in participant strategies to the ones found in Chapter 2 and 4. Indeed, the "binary switch" observed previously in our experiments, where some participants exclusively select exploratory actions, was also very pronounced across the two experimental conditions from Wu *et al.* (2018). Second, participants had a strong locality bias in their patterns of exploration.

We examined the two main claims made by Wu *et al.* (2018) in their study – specifically, participants' tendency to under-generalise and the importance of uncertainty directed search in their exploratory strategies. We looked at whether the locality bias could be explained in terms of a tendency to under-generalise about the similarity between actions. While we find that participants had a tendency to underestimate the correlation of rewards across both conditions, we found that under-generalisation could not explain alone participants' strong tendency for local exploration. Contrary to the modelling results of Wu *et al.* (2018), we also found that participants were able to adapt the extent to which they generalised to the structure of their environment. Finally, we compared our general model to the localised GP-UCB model presented by Wu *et al.* (2018). Overall, we found that our model offered a better account of participant behaviour. Contrary to their claim, we found no substantial evidence in favour of participants relying on (global) uncertainty during search. Instead we found that participants mostly relied on local search, while leveraging expected rewards and with an important drive towards novelty. One plausible hypothesis, supported by their model but not put forward clearly in their paper, is that participants aim to resolve *local* uncertainty during search. This heuristic could have several benefits - it exploits the structure of the environment efficiently while avoiding the expensive computational cost of representing the complete structure of the environment.

In this chapter, we looked at participants' decision strategies on tasks when their underlying structure is known *a priori*. We found that participants were indeed adaptive to the reward structures of the tasks in their exploratory search strategies. In Chapter 2,3 and 4 we looked at participants' ability to learn and exploit the structure of new and unknown tasks. In the next chapter, we investigate the ability of people to adapt to change and learn across tasks when their structure may vary.

Chapter 6

Garden paths and adaptive behaviour in changing environments: A resource-rational account

6.1 Introduction

At the start of this thesis, we set out to understand how people learn across a sequence of different tasks. The first learning problem people are faced with comes from the fact that the structures of our environment are not directly observable and must be inferred from our observations. To study this, we focused on people's strategies when learning across tasks that shared structural similarities. A second problem arising from the multi-contextual nature of realistic learning environments is that change can occur with or without the presence of explicit cues. This is the problem we focus on in this chapter.

When faced with changing environments, the task of the learner is to both detect that change has happened and to learn a new model of the world. To avoid restarting our learning every time we are faced with a change of context, recognising familiar patterns and generalising across contexts is essential. When adjusting to the balance of a new bicycle, facing a new opponent in a game, or selecting an item from the menu of a new restaurant, we can rely on previous experiences that will guide our actions in a way that is much more efficient than if we had to relearn everything from scratch. We can, for example, recognise patterns in someone's play that are similar to a previous match, or feel excited about seeing a dish that we tried in a different restaurant. Beneath this lies an ability to segment our experiences in a structured manner, which allows us to re-use them when relevant. How do people realise when a task shares structural similarities with a previous one? Conversely, can we detect when a context has changed and adapt to a new task without being misguided by irrelevant chunks of knowledge? For example, a recent trip to Italy might offer a misleading idea of the spaghetti dish in a Scottish pub.

This challenge of continuously learning across multiple tasks remains largely unsolved and is a topic of considerable interest (e.g. see Kirkpatrick *et al.*, 2017; Wang *et al.*, 2016; McCloskey & Cohen, 1989). Unlike humans, a difficult task for Reinforcement Learning algorithms has been to learn multiple tasks sequentially without forgetting previously acquired knowledge: a phenomenon referred to as "catastrophic forgetting". We hope that a better understanding of some of the facets of human behaviours in this domain can inform the design of better algorithms.

So far, we have looked at the strategies of participants when faced with sequences of tasks that shared the same underlying structure. We found that many participants were able to learn and exploit the underlying structure of their environment, and improved their performance across tasks. To better understand

the strategies of participants in goal directed exploration, we presented a general model of human exploration. One of the limitations we highlighted, however, was the inability to capture learning dynamics, specifically participants' ability to re-use previous knowledge and improve across tasks. Indeed, our model assumed each task to be independent. x In this chapter, we examine the human ability to self-direct their learning across multiple contexts, when the underlying problem structures may change. An environment may change gradually over time, which requires minor adaptation to the environment, however, there may also be abrupt changes that require drastic adaptation, and a revision of the structural assumptions about the environment and of the agent's behaviour (Lloyd & Leslie, 2013; Gershman *et al.*, 2015; Qian *et al.*, 2012). We investigate how the sequential nature of learning might interact with people's learning strategies. Our experiments in this chapter are designed to study to what extent participants are sensitive to changes and similarities between tasks and how this affects their learning and performance.

Studying the assumptions of participants about the world, and how people learn their expectation of change across tasks is particularly beneficial when participants are able to self-direct their learning. Indeed, participants' representation of uncertainty is a key factor for successful adaptive behaviour, as they need to rely on this when deciding to which extent they should explore more, or exploit their current state of knowledge.

When learning across different tasks, it is crucial to have an accurate model of the environment to be able to predict events and achieve desired goals. An agent's beliefs about changing contexts can be incorporated into the Bayesian perspective by specifying how the parameters (θ) representing the structure of the world evolve over time. Bayesian models can express different assumptions about the world, and hence capture different types of regularities. By focusing on, or ignoring, different parts of a learner's experience, different models allow for

contrasting predictions (Courville *et al.*, 2006). A model that assumes all tasks to be independent might behave very differently from a model that assumes them all to share the same reward structure.

Studies have shown that people are able to adapt their decision strategies in the face of change. In the case of environments where the reward structure changes gradually, participants have been shown to adapt their behaviour to the volatility of their environment similarly to ideal Bayesian agents (Speekenbrink & Konstantinidis, 2015; Behrens *et al.*, 2007; Angela, 2007). In the case of abrupt change, some studies have suggested that rather than adapting the current representation of the environment, the previous representation is abandoned altogether in favour of a new one (Bouton, 2004; Redish *et al.*, 2007).

Prior to our experiments, we hypothesised that people rely both on contextual cues to detect change (i.e. a visible change of environment) and on a learned volatility of their environment. Two sources of uncertainty thus have to be taken into account by the learner in our task. First, the possibility of change in structure of the environment, and second, the error stemming from the discrepancy between the beliefs of the learner (their current world model, or prior) and the actual structure of the world (Speekenbrink & Shanks, 2010; Qian *et al.*, 2012; Yu & Dayan, 2005; Bland & Schaefer, 2012).

We designed the sequence of tasks by using tasks that shared structural similarities (like in the previous chapters) and tasks that were fundamentally different structurally to better understand the mechanisms at play when learning in the face of change. In a first part, we present our experimental set-up before analysing the behaviour of participants. In our results, we find that participants are able to exploit the structural similarities and improve across related tasks, as well as adapt to change. In some cases however, participants were consistently unable to adapt to a new simple task. In a second part, we compare models with different

assumptions about the dynamics of change across tasks to better understand the behaviour of participants. Specifically, we explore evidence for hypothesis sampling in active learning to try to explain both when people succeed and when people fail at adapting to different tasks.

6.2 Experiment 1

In Chapter 2, we presented four experiments in which participants were presented with three grids, all sharing a common pattern. Our analysis in Chapter 2, 3 and 4 focused exclusively on these initial three grids. In this chapter we consider the full sequence of tasks presented to participants in Experiment 1, which included six further grids.

6.2.1 Methods

Participants

The experimental data in this section comes from the same experiment as Experiment 1 in Chapter 2. We recruited 79 participants using Amazon's Mechanical Turk service. They received \$0.75-\$1, which was doubled for participants whose final scores were in the top 10 percent. Following the instructions given to participants, we excluded participants whose performance was worse than chance (n = 3). We also excluded participants who failed to select more than 2 different tiles on the majority of grids (n = 5), as it showed a lack of engagement with the task.

Participants were presented with three blocks of three grids (i.e. 9 grids in total). To evaluate participants' ability to improve across blocks of grids we designed a control condition that omitted the initial block of grids. We recruited 44 further

participants for this control condition. Participants were paid according to the same reward scheme than in the experimental conditions. In the control condition, one participant was excluded for performing below chance. Two participants were excluded for failing to select more than 2 different tiles on the majority of grids.

Procedure

Participants were told they would see 9 grids composed of 81 tiles (9x9). For each grid, they had 20 turns to select tiles in order to maximise their overall score. During the experiment, the number of turns left is continuously displayed. Participants were told there may be an underlying pattern behind the reward associated to the tiles. Before seeing the first grid, participants were first presented with a familiarisation grid to learn how to select tiles, where the brightness and associated rewards are sampled independently at random. In the actual game, we use two different rules to determine the reward distributions of the grids:

1. In the **location rule**, the location of a maximum reward value tile is sampled at random from the grid. The reward associated to each tile is exponentially smaller, the further away it is from that maximum tile. We used an exponential decay constant of 0.4, yielding a large difference between the maximum and its closest neighbour, in order to reward participants that found the maximum in a grid. This was the only rule used in the experimental data presented in Chapter 2.

2. In the **brightness rule**, the rewards are linearly proportional to the brightness (i.e., rendered gray-scale values between black and white) of each tile. The brighter the tile, the higher the reward. The distribution associated with the brightness of the tiles is the same as the reward

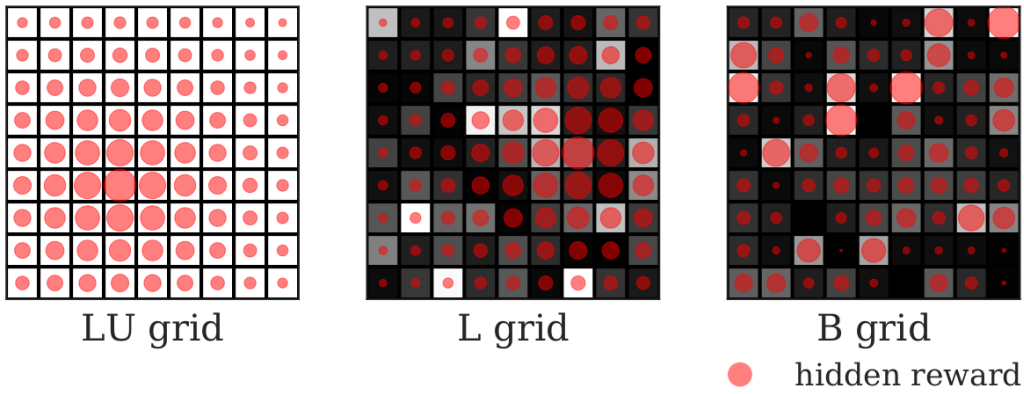


Figure 6.1: Example of each type of grid presented to participants. The size of the red circle is proportional to the reward associated with the tile. LU: Location Uniform, L: Location, B: Brightness.

distribution in the location rule, meaning that both rules share the same distribution of rewards.

To prevent participants from simply learning the maximum reward and assessing their performance with respect to that value, the maximum reward for a given grid is sampled from a normal distribution $\mathcal{N}(\mu = 200, \sigma^2 = 50^2)$. All reward values are rounded to the nearest integer in the games. We report the normalised scores (between 0 and 1) of participants to compare and evaluate their performances.

Participants were shown three blocks of three grids in orders differing according to their experimental condition. The three types of grids presented in blocks to participants are shown in Figure 6.1. The order in which the blocks were presented for each condition is detailed below and can be seen in Figure 6.2.

Our aim was to design experimental conditions to better understand people's ability to adapt to different kinds of environmental changes. Specifically, we

manipulated 1) the presence or absence of contextual cues indicating change and 2) participants' expectation that sudden change may occur.

Participants were separated into two experimental conditions: *Brightness First* and *Location First*.

In the *Brightness First* (BF) condition, participants were initially presented with a block of three grids that follow the location rule, where all tiles are of uniform brightness. We call this block **LU** for *Location rule with Uniform brightness*. Next were three grids that follow the brightness rule (or **B** grids). Finally, participants were presented with three location grids (**L** grids), where the brightness of each tile was distributed in the same way as in the B grids but was this time irrelevant to the reward function. In the BF condition we were first interested in observing whether participants can detect the change between LU and B grids, which is accompanied by a contextual cue (i.e. a different visual appearance for the different grids). Second, we were interested in their ability to detect change between B and L grids, in the absence of cues marking the change of context (B grids and L grids are visually similar).

In the second condition, named *Location First* (LF), participants were presented the three same initial LU grids, this time followed by the L grids (location rule with distracting brightness cues), and finally the three B grids. This contrasting condition was designed to first test whether participants are able to directly re-use structural knowledge between LU and L grids despite the change in appearance. Second, we wanted to observe to what extent participants would be able to detect change between L and B grids in the absence of explicit cues.

We also hypothesised that the expectation of change may have an effect of participants' ability to adapt to change. In the first BF condition, participants were

shown abrupt change with a contextual cue after 3 grids, whereas participants in the LF condition did not see any change of structure until the 7th grid.

In the control condition, participants started directly with a block of L grids. We compare their performance to the performance of LF and BF participants on their respective blocks of L grids. We do this to evaluate possible transfer effects between LU grids and L grids, which share the same underlying structure but differ in appearance. In the case of LF participants, the underlying rule remains the same but the brightness cues appear when switching to L grids. We refer to this case of knowledge re-use as *direct transfer*. We saw in Chapter 2 that participants were able to improve across LU grids. Here we look at whether participants can also detect the hidden similarity between grids, when their appearances differ (presence of brightness cues versus absence). In the BF condition, participants have to detect a change has occurred and recall the L rule after three B grids. We refer to to this as *long transfer*.

6.2.2 Results

Chapter 2, 3 and 4 focused on the performance of participants in the LU grids. Here, we start by giving a general overview of the patterns of performance in the two other blocks of grids (L and B) in both conditions to understand the general ability of people to deal with change when learning across tasks. We then conduct an analysis that examines individual differences by looking at the behaviour of participants according to the strategy types we identified in Chapter 4. We hypothesised there would be marked differences in people's ability to adapt following the type of strategies they used in the initial three grids, specifically regarding their use of a representation of the environment.

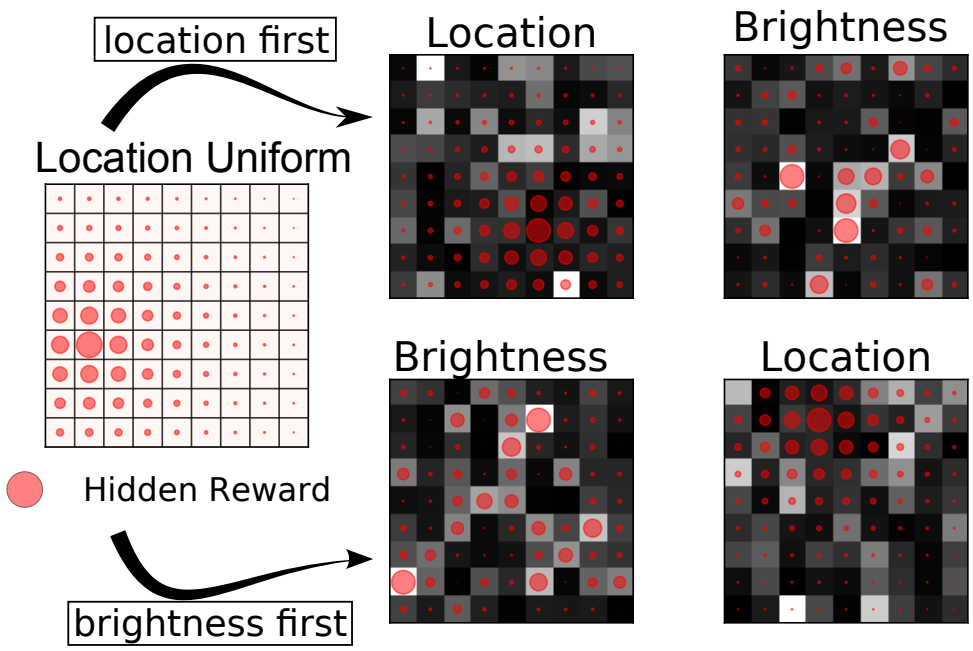


Figure 6.2: Experimental conditions in Experiment 1. The top row sequence was presented to the *Location First* condition, the bottom row sequence was shown to the *Brightness First* condition. Each grid type is shown three times in a row. Top row shows Location First condition. Bottom row shows Brightness First condition.

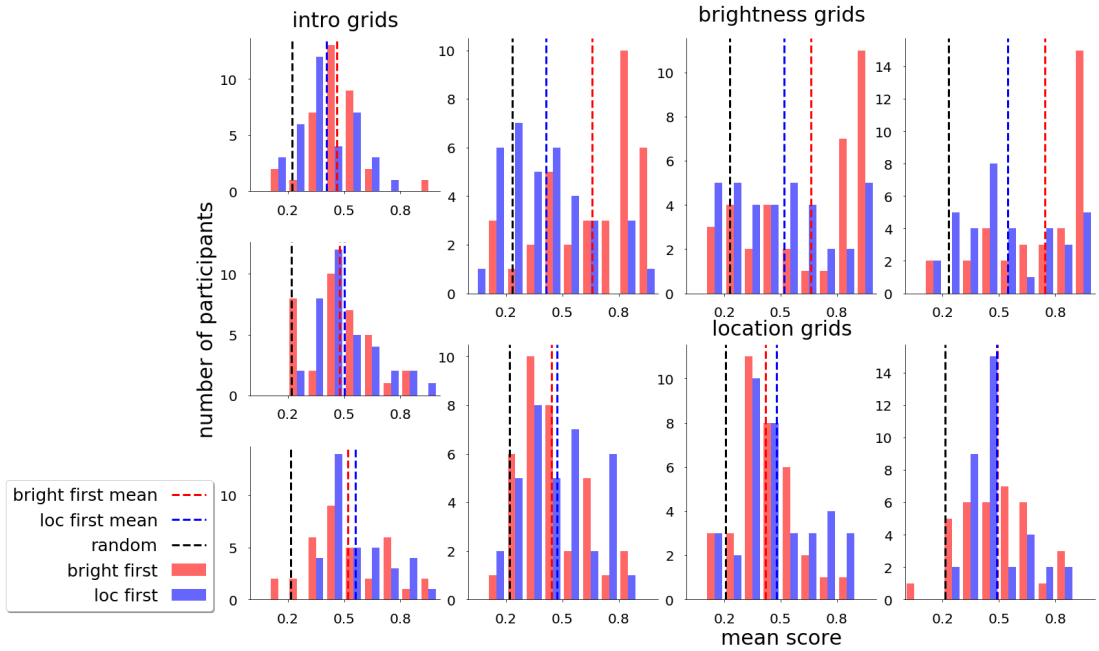


Figure 6.3: Experimental results of Location First (LF) and Brightness First (BF) conditions. Each plot contrasts the performances of participants in the LF and BF conditions. The left column shows the performances of participants on the three LU grids (intro grids). The top row shows their performances on the brightness grids. The bottom row shows their performance on the location grids. In general, participants improved across grids of the same block. BF participants had significantly better performances on the brightness grids than LF participants.

Brightness grids

In this section, we study participants' ability to adapt to a change of rule with (Brightness First condition) and without explicit cues (Location First condition). Participants in the BF condition adapted very efficiently to the brightness grids, achieving a median score of 0.79 over all three grids. This was in severe contrast with the median performance of 0.47 from the participants in the LF condition ($\Delta_{Med} = 0.32, U = 345.0, p < 0.001$) (see Figure 6.3). It's important to note that in the BF condition the rule change was accompanied by an observable change of context (from LU to B grids the brightness cues become salient), but not in the LF condition (from L to B the brightness cue distribution is similar across grids).

To assess participants' ability to improve across grids, we used a general linear model (GLM), with the reward as outcome variable. The turn and grid index were used as predictor variables. Participants in both conditions improved their performance across grids. In the BF condition, the coefficient of improvement over B grids was 0.04 ($se = 0.01, p < 0.001$), while the intercept was 0.53 ($se = 0.01, p < 0.001$). In the Location First condition, the coefficient of improvement over B grids was of 0.07 ($se = 0.01, p < 0.001$) while the intercept was 0.28 ($se = 0.02, p < 0.001$). Despite the larger grid improvement coefficient, LF participants still performed worse in their third brightness than the first brightness grid of BF participants ($\Delta_{Med} = 0.29, U = 484.0, p = 0.05$), showing a continued inability to adapt to the change in reward pattern.

Location grids

In this section, we look at participants ability to transfer structural knowledge when the tasks look visually different but shown consecutively (Location First condition), or separated by a different type of task (Brightness First condition). We

focus on participants' performance on the L grids (location rule with distracting brightness cues) in both conditions. Despite the different sequence of presentation, there was no significant difference in performance on the L grids across the two conditions ($\Delta_M = 0.03, t = 1.00, p = 0.32$). There was no significant progress in the LF condition ($b = 0.01, se = 0.01, p = 0.28$), while participants in the BF condition did show evidence for improvement ($b = 0.02, se = 0.01, p = 0.001$). This can be explained by the fact that participants in the BF condition had to detect a change of rule, while LF participants did not. To understand to what extent participants were able to re-use structural knowledge from the initial LU grids, we compare the performance of participants to that of the control condition (L grids with no pre-training on LU grids). In both conditions (LF and BF), participants had better average performances in the L grids than in the control condition. In the LF condition ($\Delta_M = 0.11, t = 3.97, p < 0.001$) this indicates that participants were able to do direct transfer, detecting that the structure was similar despite the change in appearance. In the BF condition ($\Delta_M = 0.08, t = 3.0, p = 0.003$), it indicates that participants were able to do "long transfer", and re-use the learned structure from the LU grids in the L grids. The latter shows that participants were able to detect the change of rule in the absence cues and leverage the structural similarity with LU grids.

Before presenting a summary of our analysis about people's ability to adapt to change across different contexts, we wanted to better understand the patterns of exploratory behaviour observed in participants, particularly the local bias discussed at length in the previous chapters.

Local bias in search

When analysing the behaviour of participants in the initial LU grids in Chapter 2 and 4, we found a strong bias towards selecting local actions. This locality

bias was also found in the data of Wu *et al.* (2018) we presented in Chapter 4. This local bias existed in the B grids too, both in the LF ($t = 5.47, p < 0.001$) and BF ($t = 8.55, p < 0.001$) conditions. Participants in the LF condition (M=2.71, SD=0.76) selected significantly more locally than BF participants ($\Delta_M = -0.53, t = -2.43, p = 0.02$). In the LF condition, this can be explained by the fact that participants did not adapt to the B grids and assumed the rewards were still spatially-correlated.

To better understand the nature of people's bias towards local actions, we were interested in seeing whether participants also had a "local" bias in the brightness dimension. i.e. if they were more inclined to select actions that were of similar brightness values to their previous selection. To evaluate this, we looked at the B grid selections of participants in the BF condition. This was not the case, as participants in the BF condition selected actions that were slightly more distant in brightness than random ($\Delta_M = 0.02, t = 2.06, p = 0.04$).

The fact that participants had a local bias in the B grids even when they had adapted to the change of rule and the absence of a bias in the brightness dimension both hint at the idea that it may be specific to spatial features. The locality bias could perhaps be explained by the fact that selecting on a distant tile requires more effort for participants, and thus constrains which tile they will select next. Another hypothesis could be that participants did not rely on a a local bias in the brightness dimension because the problem was uni-dimensional in the B grids, thus making it easier to explore. The L grids, on the other hand, were bi-dimensional, potentially making it more difficult to represent and explore efficiently.

Discussion of empirical results

In Chapter 2 and 3, we highlighted participants' ability to improve their performance across tasks when sequentially presented with grids of the same type. Similarly, participants were able to improve across the sequence of L and B grids. In general, we found strong evidence in support of participants being able to detect and exploit similarity between grids. This was true when the structural similarity was not directly observable, like when transitioning from LU to L grids in the Location First condition. It was also the case for non-adjacent tasks, like in the Brightness First condition, where participants were able to do long transfer, and outperformed participants in the control condition.

When looking at participants' ability to change, we found that participants were able to efficiently detect and adapt to change. Participants were able to learn and exploit the brightness rule in the Brightness First condition already from the first grid, when the change was marked by an explicit cue. Participants were also able to detect change without any explicit cues, when adapting from the B grids to the L grids in the Brightness First condition. However, participants in the Location First condition failed to adapt to the brightness rule, despite its apparent simplicity and throughout the three grids they were presented with. One hypothesis for this is that people learn an expectation of change, or a degree of volatility, about their environment. Indeed, participants in the BF condition were able to adapt both when change was observable (LU to B) and hidden (B to L). It is possible that participants learned to expect change in this case, since they experienced it early on and jointly with a visual cue marking the change of context. The inability of LF participants to adapt could be due to the assumption that the underlying pattern would remain the same, since they were presented with a sequence of six grids following the same reward pattern prior to the change of rule. We test this hypothesis with the use of computational models in Section 6.3.

In Chapter 4, we used a model of individual differences to differentiate participant strategies. In the next section we look at how participants' strategy type influenced their ability to adapt to change across different tasks.

6.2.3 Individual differences in adaptive behaviour

In the previous chapters, we highlighted the important differences that existed between participants. Specifically, we identified four groups of strategies used by participants in Chapter 4. In this section, we analyse participants' ability to adapt according to the strategy they used in the initial block of LU grids, assuming that their general patterns of behaviour would be consistent across all nine grids. We thus re-use the clusters identified in Chapter 4 and examine the performance of participants according to their respective groups. We look specifically at participants ability to adapt to brightness grids across both conditions, since the brightness rule was novel to participants. We hypothesised that participants relying on model based strategies (i.e. *Scholars* and *Maximisers*, both with dominant contributions by the GP model components) would be better at adapting than participants mostly described by the heuristics components of our general model (*Local explorers* and *Greedy locals*).

Brightness First condition

In Figure 6.4, we show the performance of participants clustered according to their strategy on LU grids. We find that the clusters had qualitatively different types of behaviours. Both Maximisers and Scholars performed better than Local explorers and Greedy locals. An analysis of variance (ANOVA) shows that there was a significant difference in performance between Local explorers and both Maximisers and Scholars ($F=4.86$, $p=0.01$). Greedy locals also performed

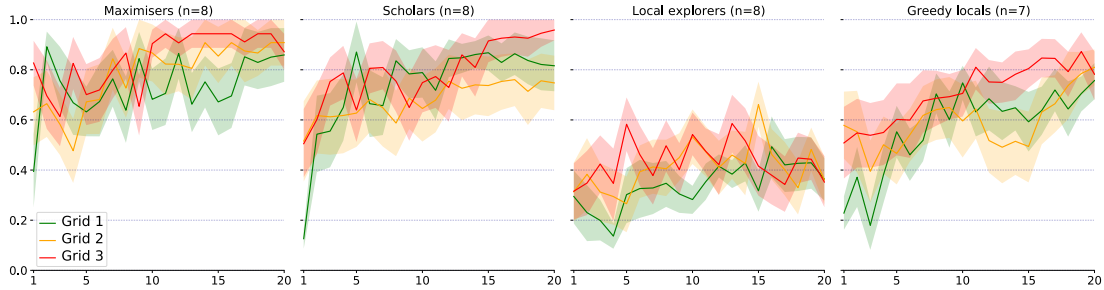


Figure 6.4: Performance of BF participants on Brightness grids according to their strategy type. The uncertainty marks the 68% Confidence Interval.

worse than Maximisers (Δ_M=0.17) and Scholars (Δ_M=0.11), though this was not significant under the Tukey post hoc test. When looking at the ratio of exploration across sub-groups, Local explorers explored significantly more than all other three groups ($F = 7.64, p < 0.001$). These results followed our hypothesis closely in that participants best fit by model based strategies (Maximisers and Scholars) adapted efficiently and had good performances, whereas participants who were best fit by heuristics (or model free) strategies adapted slower in the case of Greedy local participants, or not at all in the case of local explorers.

Location First condition

We show in Figure 6.5 the performances of the different sub-groups in the LF condition. Maximisers had the best average performance (M=0.64, SD=0.12), followed by Scholars (M=56, SD=0.21), Greedy locals (M=0.52, SD=0.17) and Local explorers (M=0.42, SD=0.08) but there were no significant differences across groups (F=2.0, p=0.14). When looking at the amount of exploration conducted by participants across the different sub-groups, Local explorers explored significantly more than Maximisers and Scholars ($F = 4.26, p = 0.01$). The difference was also substantial between Local explorers and Greedy locals (Δ_M=0.23), but

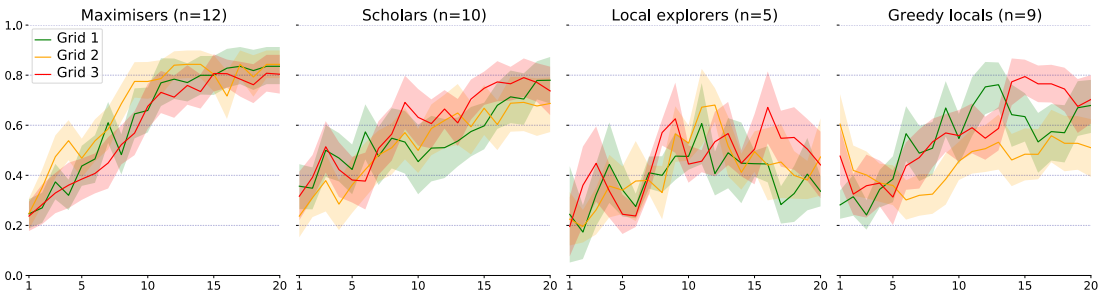


Figure 6.5: Performance of LF participants on Brightness grids according to their strategy type. The uncertainty area marks the 68% Confidence Interval.

was not significant under the Tukey post hoc test. Next, we compare the performance patterns of participants across both conditions. These results were not clearly predicted by our hypotheses, as all sub-groups failed to adapt efficiently to the change of rule. This points toward a common process that may have led participants to fail at detecting the change of context and learning the new reward structure.

Differences between Location First and Brightness First conditions

The rate of improvement for Maximisers over trials was of 0.03 (se=0.002) in the LF condition with an intercept of 0.23 (se=0.03), indicating that participants selected significantly better actions as they collected more observations. This was not the case for Maximisers in the BF condition, who had a higher intercept value (b=0.61, se=0.03) and a relatively low improvement coefficient over trials ($b = 0.01, se = 0.003$). There was no real progress across grids in the LF condition ($b = 0.01, p = 0.47$). This can also be seen in Figure 6.5, where participants show consistent linear progress for the first ten selections across all three grids, failing to select bright tiles early on. Conversely Maximisers (and Scholars) in Figure 6.4 select high value tiles from the beginning and displayed flat performance curves across trials.

Overall, we find that there were noticeable patterns of behaviour characteristic to each sub-group, though our sub-group analysis was limited by the small sample sizes. Participants in the Maximiser group showed the best performance overall, followed by Scholars and Greedy locals, while Local explorers performed consistently worst. Much like in the LU grids, Local explorers explored significantly more than the other sub-groups across both conditions. Scholars and Maximisers adapted quickly to the brightness rule in the BF condition, selecting bright tiles early on in the grid. In the LF condition, participants needed more trials to select bright tiles, and did not show any significant improvement across grids in their ability to adapt and learn the new rule.

In the next part, we use computational models to better understand participants' ability to both behave adaptively, in the case of the BF condition, and at times fail to do so dramatically, like in the case of the LF condition.

6.3 Resource rational account of adaptive behaviour

6.3.1 Detecting change

Following a Bayesian perspective, we can model an agent's representation of the environment as a generative probability distribution. We can measure the model's likelihood as the conditional probability of current observations given the model, or $p(\{\mathbf{x}, y\}|\theta)$. When an agent is faced with a new environment and does not know the structure of the task very well, it will lead to high prediction errors ($P(y_i|\theta, \{\mathbf{x}, y\}_n^{i-1})$). The prediction error will also be high if the agent knows the task structure well but a change of environment has occurred. Only

the second type of prediction error will be informative to detecting change of context. Having an accurate model of the environment is therefore essential to detecting change. It is particularly important to decide what observations to take into consideration when making predictions about the world is, as it is essential that the representation of the environment reflects the current state of the world accurately and supports the prediction of future states. We will refer to the uncertainty about the underlying structure of a task as *structural uncertainty*. High prediction errors have been shown to positively correlate to the learning rate of participants (Behrens *et al.*, 2007; Nassar *et al.*, 2010; Speekenbrink & Konstantinidis, 2015; Courville *et al.*, 2006). This indicates that people might adapt and learn in changing contexts in an optimal (or near optimal) way: fine-tuning their model with a small learning rate in the case of stable environments, and revising their beliefs aggressively when their prediction error is high.

Intuitively, having an accurate expectation about change in the environment will benefit the learner, particularly in situations where there is ambiguity about change of context. Ambiguity can arise when the underlying structure is uncertain because the task is difficult, or because only few data have been observed. If a learner assumes that an environment is stable, they are unlikely to be aware of the context ambiguity and will fail to detect and adapt to a new context. Conversely, if a learner assumes that change is very likely, they may fail to generalise and overfit the data (O'Reilly, 2013).

Ideally, the learner would have a correct estimation of the probability of change before the learning begins, but this is only possible when there is some familiarity with the task environment. When faced with a new environment with cues and features different from anything that's been encountered before this is not possible, since there is no way of knowing when or how the environment will change.

In the next section, we present a family of algorithms as a suitable candidate to

explain how people might update their beliefs about the underlying structure of tasks in a changing world despite these difficulties.

6.3.2 Inference by sampling

Approximate inference methods, that focus on managing the trade-off between computation time with accuracy, have offered an interesting ground for theories that take into account cognitive constraints. One particularly popular theory comes from sampling algorithms. These algorithms approximate complex probabilistic distributions by using a collection of samples (Sanborn, 2017). The idea that each individual might only hold one of few samples from the posterior distribution over hypotheses has been a recently popular model of cognitive constraints, serving as a bridge between rational analysis and process models (Griffiths *et al.*, 2012; Goodman *et al.*, 2008; Vul *et al.*, 2014). These kind of models, named rational process models, have been used to successfully account for a range of cognitive biases that Bayesian models had left unexplained (Sanborn *et al.*, 2010).

One important signature of inference by sampling (also known as 'hypothesis sampling') is sensitivity to the order of data. Particle filters is an example of such hypothesis sampling models. Hypothesis sampling refers to the idea that people maintain a tractable number of individual hypotheses (or 'particles') instead of having a representation of the complete posterior. Under a particle filter model, a learner can trade off cognitive resources with accuracy by limiting the number of samples. When considering a smaller set of hypotheses, a learner is able to reduce the cognitive resources necessary (i.e. memory and computation), at the cost of a less accurate approximation of the posterior. This family of algorithms has been recently used to produce order effects consistent with human behaviour across a range of tasks, such as causal learning (Abbott *et al.*, 2011), category

learning (Sanborn *et al.*, 2006) or symbolic concept learning (Thaker *et al.*, 2017) that were otherwise unexplained in a Bayesian framework.

We thus turn to particle filter to understand whether the sequential effects observed in our experimental data can be explained by the behaviour of particle filter algorithms. In the next section, we introduce a general template for the particle filter algorithm as well as the the parameters that influence its behaviour.

6.3.3 Particle Filters

Particle filtering is an algorithm that belongs to the family of sequential Monte-Carlo methods (SMC) (Doucet & Johansen, 2009). One difficulty in MCMC methods is to sample from the posterior distribution. A frequent way of overcoming this, is to sample a set of particles $\left\{h^1, \ldots, h^M\right\}$ from a proposal distribution (e.g. the prior $Q(h)$) and weight those to make up for the fact that they were not direct samples from the posterior.

$$h^m \sim Q(h), \quad w^m \propto \frac{P\left(X|h^m\right) P\left(h^m\right)}{Q\left(h^m\right)}$$

The weighted particles are then used as the approximation to the posterior. It is often the case, that some or most hypotheses are irrelevant to the posterior, and only a few particles will be assigned all of the probability weight. This is a problem known as the *degeneracy* of the sample set. To rid of this, we can delete the unlikely hypotheses, and resample from (or close to) the important hypotheses, according to their importance weights. During the resampling step, a new set of particles are proposed and selected from a distribution based on the previous set of particles and their normalised importance weights.

$$\tilde{h}_n^m \sim P\left(h_n|h_{n-1}\right), \quad w_n^m \propto w_{n-1}^m \frac{P\left(x_n|h^m\right) P\left(h^m\right)}{Q\left(h^m\right)}$$

Particle filtering solves the problem of having to recompute the posterior over all previous data points at every new observation by updating the particle weights online. At each time step, the posterior is approximated through the discrete distribution of each particle and its normalised weight. In this case, we take $Q(h)$ to be the prior $P(h)$, which is easy to sample from. This simplifies the weights to $w_n^m \propto P\left(x_n|h^m\right)$, i.e. the normalised likelihoods.

We investigate the explanatory power of hypothesis sampling by looking at model simulations. Participants select a sequence of actions and have to evaluate possible hypotheses about the underlying reward structure of the current grid given their observations. Under our model a hypothesis is represented by a particle h^m, which consists of the hyperparameters of a Gaussian Process model. These hyperparameters define the importance of each feature (the *x,y* coordinates of a tile and its brightness b) and their functional relationship to a tile's associated reward. In other words, these parameters (jointly with the choice of the kernel) define the space of functions considered by the GP. The particle filter is responsible for the inference problem of estimating the GP parameters.

Like we did for the general model, we use an RBF kernel as a generative model over possible functions (see Section 3.3.2). For simplicity, we use Thompson sampling as a decision strategy. Thompson sampling, unlike the Upper Confidence Bound acquisition function, is parameter free and corresponds to probability matching – a pattern of behaviour observed in people (see Vul *et al.*, 2014, for a review). To trade off between exploration and exploitation, Thompson sampling draws a sample from the posterior (provided by the GP in our model) and selects the best action given that sample. Sampling from the posterior takes into account both the

expected value of actions and their associated uncertainty. This provides a balance between the exploration of uncertain actions and selecting known rewarding ones.

6.4 Model simulations

6.4.1 Particle Filter parameters

To understand the ability of participants to progress across grids, we first look at the importance of the initial set of particles. We consider an initial proposal that is appropriate to the task structure (*correct prior*). For this we sample the initial particles from a half-normal distribution with the mean set close to the empirical parameters (SD=0.05). We sample the length-scales from a gamma distribution centred around their empirical parameters (shape=3, scale=.15). We contrast it with a proposal where the particles do not correspond to the grids' reward structures (*incorrect prior*). We use a non-informative (uniform) prior for the GP kernel weight parameters and an exponential distribution (μ=0.01) for the length-scales, implying a low correlation between features and rewards. Given these two priors, we compare three resampling schemes:

No Resampling In this scheme, the initial set of particles is carried over without ever resampling. This means that the set of hypotheses considered remains constant throughout the different tasks. This scheme can be understood as treating all tasks as independent of one another given that no learning happens across tasks since re-weighting happens at every time step independently.

Static Adaptive Resampling Here, particles are resampled when the variance

of the weights is too large, i.e. when only a few particles carry explanatory power. This is calculated by the Effective Sample Size (ESS), where the ESS $\approx \|\mathbf{w}_t\|^{-2}$. Similarly to Abbott *et al.* (2011), we set a threshold at 0.10N, ten percent of the number of particles. Successful particles are carried over according to their weights following the systematic resampling algorithm (Doucet & Johansen, 2009).

Adaptive Resampling with jitter This last resampling scheme uses the same ESS threshold as the previous model and resampling algorithm. During resampling, new particles are sampled from a Cauchy transition kernel (γ=0.1) centred around the resampled particle. The Cauchy distribution has most of its probability mass narrowly distributed around 0, but has heavy tails, inducing occasional rare "jumps" in parameter values. While this disturbance introduces imprecisions in the approximation, these are just added noise and allow the introduction of new hypotheses to the particle population. Introducing new particles to the population is often referred to as a "rejuvenation step" and leads to considering a more diverse set of hypotheses. We name this sampling scheme "Adaptive Resampling with jitter" to not confuse it with rejuvenation schemes that rely on Metropolis-Hastings steps. Abbott *et al.* (2011) suggested that rejuvenation and resampling could correspond to deliberative reasoning, a process that is more computationally expensive than simply updating the weights of particles. Considering new or alternative hypotheses adaptively could be a resource rational strategy, triggered by a given state of the environment or of the learner, saving them from constantly having to evaluate an intractable amount of hypotheses at every time step.

6.4.2 Explaining transfer and adaptation across tasks

In this section, we look at simulations of particle filter models with different assumptions about learning dynamics to better understand the behavioural phenomena observed in our experiments. We first compare the performance results of particle filter models on the Brightness First condition. We focus on the phenomena of direct transfer and adaptation to change, and omit the last block of L grids. We discuss the case of "long transfer" in a later section.

We show the performance of BF participants on the first 6 grids in Figure 6.6. Here, we look specifically at the performance of participants who showed they were able to exploit the structure of the grids by reselecting tiles in the tasks (n=23) (named *explore-exploit* participants in Chapter 2). The two phenomena we are interested in capturing here are participants' ability to progress across grids, and the ability to adapt to the change of reward structure.

Of the six models we evaluated, the model with the *adaptive resampling with jitter* scheme and an incorrect initial proposal showed both progress across grids and the ability to adapt to the change of rule (see Figure 6.7). Models with an appropriate proposal for the initial set of particles did not show progress across grids since they already showed good performances in the initial grid. Models with *static adaptive resampling* did not adapt to the change of rule, as the particles coherent with the B grids were filtered out during the LU grids. The *never resample* scheme adapted to the change of rule, but did not progress across grids.

Here, the particle filter account of participant behaviour presents transfer across similar tasks as a gradually more accurate representation of the posterior that happens by resampling around the best particles from previous grids. Adapting to a change of task structure happens through deliberative reasoning, which happens strategically when the hypotheses under consideration explain the observed data

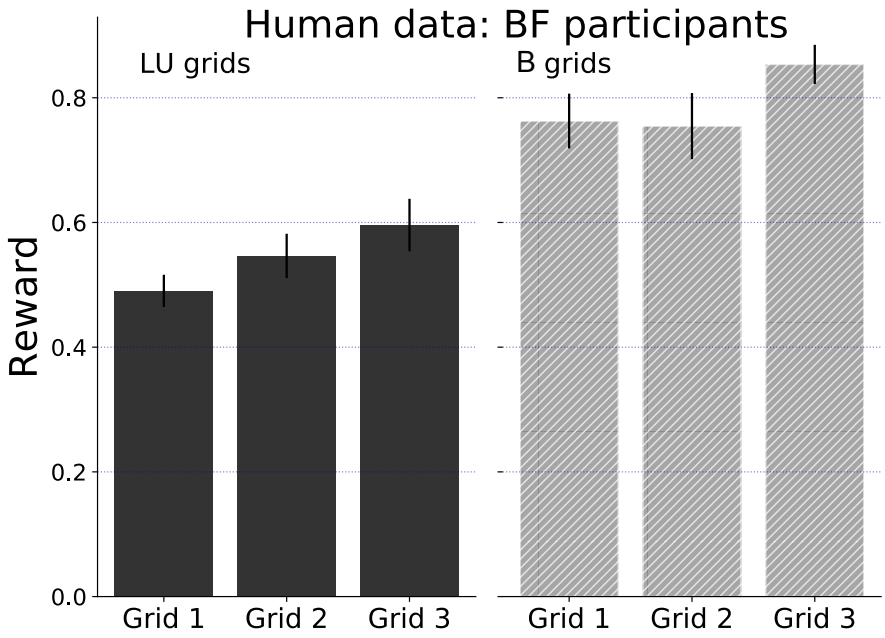


Figure 6.6: Human data showing participants' ability to progress across trials and adapt to a change of environment structure. The error bars show the SEM.

poorly. Next, we focus on participants' inability to adapt to the change of rule in the LF condition and whether it can be explained by hypothesis sampling.

6.4.3 Garden paths in self-directed learning

In this section we consider participants' inability to adapt to a change of task. The performance of LF participants was significantly worse on the B grids than BF participants (see Figure 6.8), and LF participants failed to adapt throughout all three grids.

Again, we consider the six models presented earlier. The only model that performed well on the L tasks, but did not adapt to the B grids was the particle filter with *static adaptive resampling* and with a correct initial proposal distribution. These simulation results suggest that hypothesis sampling can

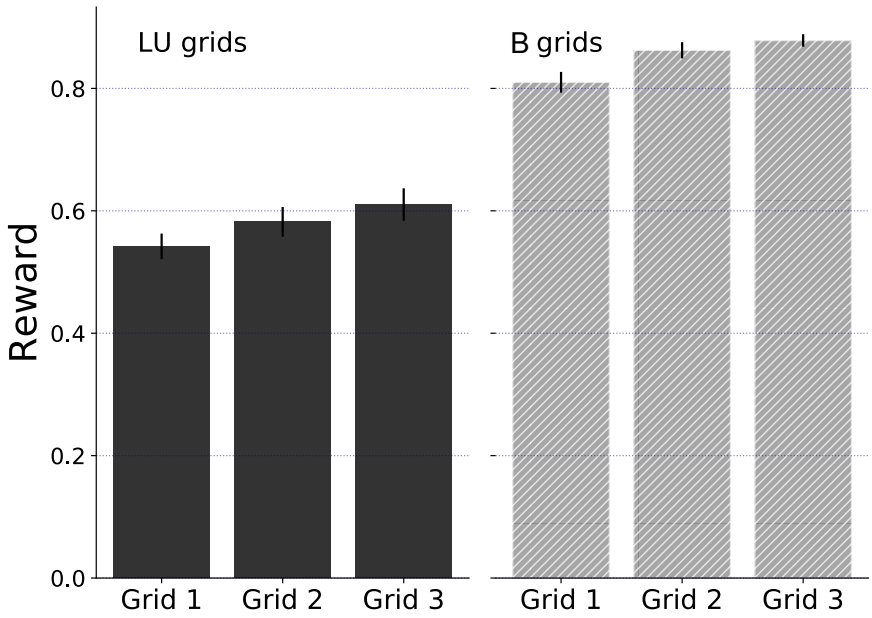


Figure 6.7: Performance of particle filter model with adaptive resampling with jitter, and an incorrect proposal for the initial set of particles. The model progressed across trials, and was able to adapt from the Location reward pattern to the Brightness reward pattern. The error bars show the standard error of the mean.

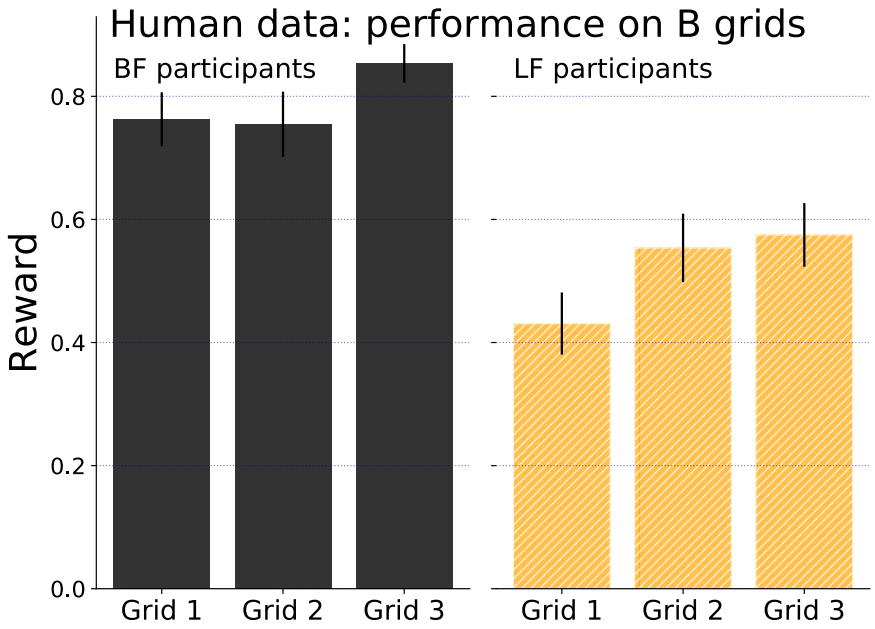


Figure 6.8: Human data: Performance of Brightness First and Location First participants on the three Brightness grids.

explain the poor performance of participants as a garden path, where participants were not able to consider alternative hypotheses than the ones considered during the previous grids. In Figure 6.9, we contrast the performance of the two particle filter models that best matched participant behaviour in the BF and LF conditions.

In the previous section, we found that resampling with jitter could explain both transfer across tasks of similar nature, and the ability of participants to adapt to change when the structure was new. In the LF condition, static resampling (i.e. no renewal of the particle set) explained participants' inability to adapt. What could trigger participants to rely on a strategy akin to resampling with jitter in one case but not the other? We suggest two plausible explanations: 1) Participants did not realise the environment had changed, or 2) they did, but failed to generate better hypotheses during resampling. If the former is true, it would support deliberative reasoning being triggered by a state of the agent, and not just their environment.

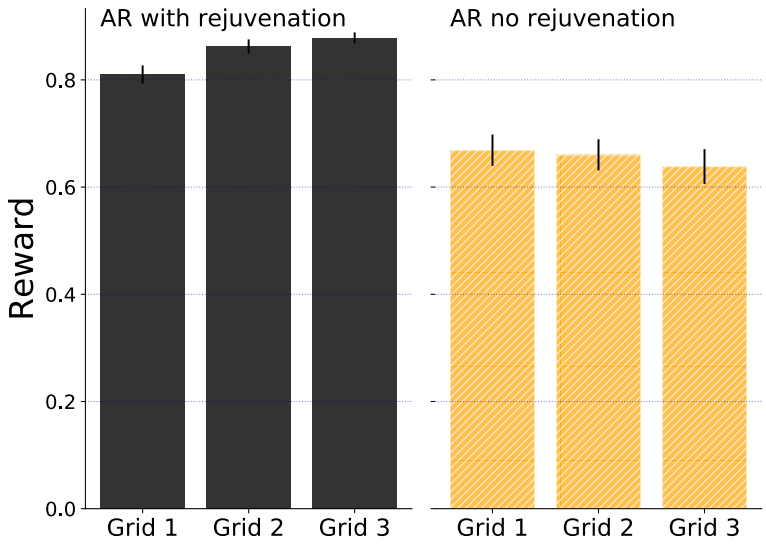


Figure 6.9: Performance of particle filter models best matching participant behaviour in the B grids. In the BF condition (black) the best matching model was a particle filter with adaptive resampling with jitter. In the LF condition (yellow) it was a particle filter with Static Adaptive Resampling .

It is possible, for example, that people rely on their expectation of change, i.e. an estimation of the volatility of their environment, to inform their decision to spend mental resources on generating alternative hypotheses. In the LF condition, participants would have learned the environment has very low volatility after six grids following the L rule, and thus failed to consider the environment may have changed.

Some studies have highlighted that high prediction errors positively correlate to the learning rate of participants (Behrens *et al.*, 2007; Nassar *et al.*, 2010; Speekenbrink & Konstantinidis, 2015; Courville *et al.*, 2006). This indicates that people might adapt and learn in changing contexts in an optimal (or near optimal) way: fine-tuning their model with a small learning rate in the case of stable environments, and revising their beliefs aggressively when their prediction error is high. However, because of the change of reward structure, we expect participants to have had relatively high prediction errors when changing to the B grids. This leads us to consider the second hypothesis, namely that participants did not find a fitting hypothesis after realising the reward structure had changed. This is perhaps surprising given the simplicity of the Brightness rule, and the high performance of participants in the BF condition. Participants' inability to generate an appropriate hypothesis could be explained by a phenomenon known as *dimensionally selective attention*. A number of studies have shown that people learn to selectively attend relevant dimensions (Niv *et al.*, 2015; Kruschke, 1992; Nosofsky, 1986). Selective attention can be linked to assuming the number of relevant features to be sparse when learning the structure of the world (Wilson & Niv, 2011; Gershman *et al.*, 2010a; Kemp & Tenenbaum, 2009). This sparsity assumption has been shown to greatly alleviate the computational cost of the RL problem. In our experiment, selective attention could have caused people to exclusively consider hypotheses in the x,y dimensions, and thus fail to consider

hypotheses involving the brightness dimension since it was irrelevant in previous grids.

6.4.4 Representing structure, long transfer and memory

Though we did not examine long transfer in our models, we discuss here how memory mechanisms might support participants' ability to re-use knowledge across non-adjacent tasks. In the BF condition, we saw that participants were able to improve their performance on the L grids thanks to having learned the location rule in the LU grids earlier. This was the case despite no explicit cueing that change had happened (B and L grids were visually similar) *and* LU and L grids being visually distinct. It implies that participants were able to learn the hidden reward structure separate from the visual context of the task.

In our particle filter model, a context θ_i is represented by a set of particles and their associated weights. Our model considered all contextual features, implying that participants would conduct inference over the complete space of hypotheses given those features. Our discussion regarding selective attention suggests that a fixed model that comprises all the different contextual features of the environment is unlikely to be an accurate model of people's representations. Indeed, adapting to a change of context may involve more complex mechanisms than simply updating the parameters of an exhaustive generative model. A more likely theory is that people create lower-dimensional representations of their environment with features that are considered relevant for the prediction of rewards, and only reconsider the causal structure of their environment when judged necessary. People may then hold a bag-of-contexts with the structures of the previously learned tasks. Naively, this could be implemented in the particle filter template by keeping particles in store across tasks, but it would not explain how people are able to parse experiences as novel or similar to previous

ones. Without a structuring of past experiences, an infinite amount of world representations would need to be stored. Bayesian models have been suggested to explain how people may decide what should be considered a new context vs one that belongs to a known category (Niv, 2019; Qian *et al.*, 2012; Gershman *et al.*, 2010b). In their accounts, experiences (or observations) are clustered according to the causal models that generate them, and new causal models are created following an Infinite-capacity mixture model, that infers the number of clusters based on previous observations. By maximising both the within-cluster similarity and the inter-context difference, the number of different contexts can be inferred directly from the data. To control for overfitting, and limit the number of models kept in memory, one can place a prior that favours fewer contexts. A similar account was also discussed in the case of category learning by Sanborn *et al.* (2006).

6.5 Conclusion

In this chapter, we looked at participants' ability to learn across a sequence of tasks in which the underlying task structures may change. We designed an experiment to examine different aspects of learning in a changing world. In Chapter 2, we discussed participants' capacity to transfer knowledge across related and adjacent tasks and improve their performance. In this chapter, we also looked at the ability to adapt to change in the presence of an explicit cue, and *without* any explicit cues. Finally, we looked at "long transfer", or the ability to re-use an abstract task structure in a non adjacent task. Our experimental results showed that participants were able to do all these things. However, we also found that participants were in one condition unable to detect and adapt to a simple change of reward structure - and this throughout three successive grids. To better understand the behaviour of participants, we explored

"hypothesis sampling" to explain both the successes and failures they showed on our tasks. Our particle filter models showed that transfer across sequential tasks could be explained as a gradual improvement of the posterior approximation representing the task structure, and the re-use of particles from one task to the next. Adaptive resampling with jitter explained participants' ability to efficiently adapt to a change of task. Conversely, adaptive resampling without considering new hypotheses led to strong garden path effects, as appropriate hypotheses were filtered out early. This predicted participants' inability to detect change and adapt to a new reward structure. We discussed that the introduction of new hypotheses set may be a strategic tool used by participants, triggered not only by their environment but by their internal state. We suggested learned environment volatility as a potential mechanism. Participants' continued inability to consider a fitting hypothesis could also have been caused by selective attention, or the preference for lower dimensional representations.

Overall, this leaves us with a set of questions for future research. First, how are processes such as selective attention implemented trial-by-trial? Future experiments could examine how the amount of trials in L grids affect participants' ability to adapt to the change of reward structure. Would participants have been able to adapt had they only had one L grid instead of three?

Another question is whether simpler task structures (e.g. a location rule in only one dimension) would facilitate considering alternative hypotheses. It could be that the complexity of the location rule made it difficult to definitely rule out the possibility that x and y were predictors for rewards when switching to the B grids. Future experiments could investigate how people are able to disentangle different types of uncertainty during self-directed learning.

Finally, we discussed that deliberative reasoning was likely not triggered only by the environment, but also by the internal state of the learner. We suggested a

learned volatility of the environment as a potential mechanism that would induce the consideration of alternative hypotheses. When modelling order effects in causal learning, Abbott *et al.* (2011) used data from a study conducted by Collins & Shanks (2002). In their study, they found that the responses of participants, and hence their inferences, were influenced by the frequency at which they were asked to give their judgements. Abbott *et al.* (2011) suggest that such prompts could trigger deliberative reasoning. Designing interventions could help better understand the deliberative process of participants and how it may be used strategically.

Chapter 7

Conclusions

I started this thesis by posing the general question of how people select actions in order to jointly learn about the world and achieve goals within it when faced with sequences of tasks. Through experimental work and the use of computational models, I investigated how people direct their learning and select actions across tasks that may or may not share structural similarities. While the empirical evidence collected focused on the mechanisms underlying human abilities – including learning from sparse data, transferring knowledge across tasks, and the ability to detect change – still beyond those of conventional machine learning algorithms, the behaviour of participants in our experimental tasks also pointed at the cognitive constraints that may influence people's active learning strategies. Before discussing the implications of this work, the questions it raises, and directions for future work, I will recapitulate the results and conclusions drawn from the previous chapters.

Conclusion 1: There are meaningful differences in people's exploratory strategies

The experiments reported in Chapter 2 and the subsequent model-based

analyses conducted in Chapter 4 showed there were consistent patterns of variation in people's decision strategies. One striking example of this was the drive to explore of some participants. In three of the four experimental conditions presented in Chapter 2, a significant proportion of participants dismissed reward incentives and engaged in exclusively exploratory behaviour, never re-selecting actions known to be highly rewarding. This behaviour was also observed in the experimental data of Wu *et al.* (2018) that I analysed in Chapter 5. In Chapter 4 and 5, I used a model of individual differences to get a richer description of the different strategies used by participants. One notable pattern in differences amongst participants was their use of model based vs heuristics strategies. While some were able to construct representations that helped them progress across similar tasks, others relied on cheaper strategies largely agnostic to the underlying structure of their environment. Chapter 6 showed that when participants constructed an internal representation of their environment early on, it could predict their ability to adapt to a change of task.

Conclusion 2: People favour local search over globally informative actions

Important research has been conducted to understand the mechanisms behind human exploration, and studies have shown that measures of uncertainty should, rationally, be expected to be main drivers of exploration (Cohen *et al.*, 2007). Recent accounts of human exploration have characterised it as a combination of random and uncertainty directed exploration (Schulz & Gershman, 2019; Wilson *et al.*, 2014; Wu *et al.*, 2018). Our empirical analyses presented in Chapter 2 and 5 suggest that rather than seeking to reduce "global uncertainty", or maximising information gain about the structure of the environment, participants explored locally, selecting actions close to previous ones. This could be explained as a heuristic strategy, akin to hill-climbing, where participants seek to learn about the local gradient rather than constructing a complete model of their environment.

This account of human search, namely the preference for local uncertainty over global uncertainty, is coherent with results in causal learning that showed participants preferring actions that resolve uncertainty about few hypotheses rather than many (Markant *et al.*, 2016b), or seek evidence to guide local updates of their model of the world rather than resolving overall uncertainty (Bramley *et al.*, 2017).

Conclusion 3: People's exploratory strategies are adaptive to their environment

Across four experiments in Chapter 2, I sought to understand the strong exploratory drive of a significant group of participants. I found two cooperating factors underlying people's epistemic drive. First, people were motivated to explore in order to reduce uncertainty: Fewer participants engaged in exclusively exploratory behaviour when they were previously trained on the task structure. Second, and perhaps more surprisingly, people displayed novelty-seeking behaviour, or motivation to observe new evidence regardless of its informativeness. This was modulated by the memory demands of the tasks: When previous observations remained visible, more participants selected exclusively novel actions throughout. It was only when participants were both familiar with the task *and* previous observations did not remain available that this entirely exploratory strategy was not observed, and all participants used strategies that traded off between exploration and exploitation.

In Chapter 5, I looked at people's ability to generalise to guide their search process and exploit the hidden structure of their environment. I found that the degree to which participants generalised was adaptive to the correlation structure of rewards: They correctly assumed a higher degree of correlation when the structure of rewards was smooth, and a lower degree when the structure was rough.

Conclusion 4: The design of cognitive models should emphasise posterior predictive checks

In Chapter 3, I introduced a general modelling framework to study participants' strategies. I studied model simulations generated from the parameters fit to participants. I compared these simulations to the actual behaviour of participants to better understand the qualitative factors our models were able to capture, and the ones that were missing from the model. In Chapter 5, I also contrasted the predictions of our model against the one presented by Wu *et al.* (2018). In both cases, model checking, or "posterior predictive checks" (Gelman & Shalizi, 2013), was of considerable importance – both for the interpretation of the semantics of a cognitive model and for informing the design of better models.

Conclusion 5: Hypothesis sampling can explain people's successes and failures when adapting to changing environments

In Chapter 6, I presented empirical evidence for people's ability to learn in environments where tasks may or may not share structural similarity, and unexpected change may occur. I showed that hypothesis sampling could help explain distinct phenomena relating to the dynamics of learning across tasks. Our models were able to explain people's ability to progress across tasks when they shared structural similarities, their ability to adapt to change, but also specific contexts where participants were continuously unable to realise the world had changed. I discussed deliberative reasoning as a strategic tool used by participants, triggered not only by their environment but also by their internal state. My analysis suggested that other mechanisms might be at play during learning in complex and changing environments, such as learning the volatility of the environment, selective attention, and the preference for lower dimensional representations.

7.1 Future directions

Exploration and search are essential processes that guide learning, both when seeking information in the world or searching for hypotheses in the mind. Two questions arise naturally, stemming from the work presented here. First, how might one extend the models presented in Chapter 3 and Chapter 6? Second, how does the empirical phenomena presented here connect to human behaviour beyond grid tasks?

In Chapter 3, I discussed that while our general model had the benefit of offering a shared parameter space to study differences amongst individuals and offer informative descriptions of participant behaviours, it could not capture specific strategies such as the line-search heuristic. I also showed that the general model was unable to deal with learning dynamics, or change of strategies between tasks. Future models of human active learning in sequences of tasks should consider how people adaptively select strategies, and trade off between the cost of computation, expected rewards, and the structure of their environment (e.g. see Lieder *et al.*, 2014). The questions of how people learn active learning strategies and their developmental trajectories are also of considerable interest.

In Chapter 6, I considered the problem of how participants adapt to change, and looked at the dynamics of learning in sequences of tasks. While I found that hypothesis sampling was a promising algorithmic theory for how people learn and update their beliefs about the world, I also listed a number of open questions regarding the representations people might hold as they learn. How do people tease apart different kinds of uncertainty in their representations of their environments? How do people learn an expectation of change trial-by-trial? How do people learn which features to attend in new environments, and how does this process interact with the construction of task representations? Beyond the representations of single tasks, how do people segment previous experiences in

their memory to then be able to re-use them as building blocks when faced with new tasks?

In this thesis, I used grid tasks as abstractions of many real world problems with vast decision spaces. A number of recent studies have pointed that the brain may organise spatial and non-spatial information by using similar representations (Constantinescu *et al.*, 2016; Garvert *et al.*, 2017; Kaplan *et al.*, 2017). Similarly, there is consistent behavioural evidence for generalised cognitive search processes (Hills *et al.*, 2008). Future research could look into the link between how people search in the world and how people search for hypotheses.

The theoretical advances made in active learning and exploration have led to studies of people's exploratory strategies in the real world. For example, Murdock *et al.* (2017) used the notebooks of Charles Darwin to understand his reading patterns and identified shifts in phases of exploration and exploration. In a more contemporary fashion, Schulz *et al.* (2018a) looked at the decision strategies of people when making food orders using a large online food delivery service and found patterns consistent with the behaviour of participants laboratory tasks. Exploration is an essential part of every individual's journey of lifelong learning. The vast amount of information available today and the often ambiguous interests and goals of individuals make the process of searching for information increasingly complex. With a better understanding of how people represent their environments during exploration, and the strategies they follow, Machine Learning methods could improve the design of interfaces to match more natural representations, and support intuitive strategies for information gathering.

Bibliography

Abbott, J.T., Griffiths, T.L. *et al.* (2011). Exploring the influence of particle filter parameters on order effects in causal learning. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2950–2955.

Acerbi, L. & Ji, W. (2017). Practical bayesian optimization for model fitting with bayesian adaptive direct search. In *Advances in Neural Information Processing Systems*, 1836–1846.

Acuna, D. & Schrater, P.R. (2009). Structure learning in human sequential decision-making. In *Advances in Neural Information Processing Systems*, 1–8.

Addicott, M., Pearson, J., Sweitzer, M., Barack, D. & Platt, M. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, **42**, 1931.

Addicott, M.A., Pearson, J.M., Wilson, J., Platt, M.L. & McClernon, F.J. (2013). Smoking and the bandit: A preliminary study of smoker and nonsmoker differences in exploratory behavior measured with a multiarmed bandit task. *Experimental and Clinical Psychopharmacology*, **21**, 66.

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, 215–222, Springer.

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, **14**, 471–485.

Angela, J.Y. (2007). Adaptive behavior: Humans act as bayesian learners. *Current Biology*, **17**, R977–R980.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, 370–418.

Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. (2007). Learning the value of information in an uncertain world. *Nature neuroscience*, **10**, 1214.

Bernardo, J.M. & Smith, A.F. (2009). *Bayesian theory*, vol. 405. John Wiley & Sons.

Bland, A.R. & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in neuroscience*, **6**.

Bonawitz, E., Denison, S., Gopnik, A. & Griffiths, T.L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, **74**, 35–65.

Borji, A. & Itti, L. (2013). Bayesian optimization explains human active search. In *Advances in neural information processing systems*, 55–63.

Bouton, M.E. (2004). Context and behavioral processes in extinction. *Learning & memory*, **11**, 485–494.

Box, G.E. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, 791–799.

Bramley, N.R., Lagnado, D.A. & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **41**, 708.

Bramley, N.R., Dayan, P., Griffiths, T.L. & Lagnado, D.A. (2017). Formalizing neurath?s ship: Approximate algorithms for online causal learning. *Psychological review*, **124**, 301.

Brehmer, B. (1976). Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psychology*, **17**, 309–312.

Bruner, J.S. (1961). The act of discovery. *Harvard educational review*.

Busemeyer, J.R., Byun, E., Delosh, E.L. & McDaniel, M.A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks.

Chapelle, O. & Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

Clark, L., Averbeck, B., Payer, D., Sescousse, G., Winstanley, C.A. & Xue, G. (2013). Pathological choice: the neuroscience of gambling and gambling addiction. *Journal of Neuroscience*, **33**, 17617–17623.

Coenen, A., Nelson, J.D. & Gureckis, T. (2017). Asking the right questions about human inquiry.

Cohen, J.D., McClure, S.M. & Yu, A.J. (2007). Should i stay or should i go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **362**, 933–942.

Collins, A.G. & Frank, M.J. (2014). Opponent actor learning (opal): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological review*, **121**, 337.

Collins, D.J. & Shanks, D.R. (2002). Momentary and integrative response strategies in causal judgment. *Memory & Cognition*, **30**, 1138–1147.

Constantinescu, A.O., O'Reilly, J.X. & Behrens, T.E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, **352**, 1464–1468.

Cook, C., Goodman, N.D. & Schulz, L.E. (2011). Where science starts: Spontaneous experiments in preschoolers? exploratory play. *Cognition*, **120**, 341–349.

Courville, A.C., Daw, N.D. & Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, **10**, 294–300.

Daw, N. & Courville, A. (2008). The pigeon as particle filter. *Advances in neural information processing systems*, **20**, 369–376.

Daw, N.D., O'doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876.

Djamshidian, A., O'sullivan, S.S., Wittmann, B.C., Lees, A.J. & Averbeck, B.B. (2011). Novelty seeking behaviour in parkinson's disease. *Neuropsychologia*, **49**, 2483–2488.

Doucet, A. & Johansen, A.M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, **12**, 3.

Freeman, S., Eddy, S.L., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H. & Wenderoth, M.P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, **111**, 8410–8415.

Garvert, M.M., Dolan, R.J. & Behrens, T.E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *Elife*, **6**, e17086.

Gelman, A. & Shalizi, C.R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, **66**, 8–38.

Gershman, S., Cohen, J. & Niv, Y. (2010a). Learning to selectively attend. In *Proceedings of the Cognitive Science Society*, vol. 32.

Gershman, S.J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, **173**, 34–42.

Gershman, S.J. & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in cognitive science*, **7**, 391–415.

Gershman, S.J., Blei, D.M. & Niv, Y. (2010b). Context, learning, and extinction. *Psychological review*, **117**, 197.

Gershman, S.J., Norman, K.A. & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, **5**, 43–50.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on psychological science*, **3**, 20–29.

Goodman, N.D., Tenenbaum, J.B., Feldman, J. & Griffiths, T.L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, **32**, 108–154.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, **111**, 3.

Gottlieb, J., Oudeyer, P.Y., Lopes, M. & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, **17**, 585–593.

Griffiths, T., Canini, K., Sanborn, A. & Navarro, D. (2007a). Unifying rational models of categorization via the hierarchical dirichlet process.

Griffiths, T.L., Steyvers, M. & Tenenbaum, J.B. (2007b). Topics in semantic representation. *Psychological review*, **114**, 211.

Griffiths, T.L., Vul, E. & Sanborn, A.N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, **21**, 263–268.

Griffiths, T.L., Lieder, F. & Goodman, N.D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, **7**, 217–229.

Gureckis, T.M. & Markant, D. (2009). Active learning strategies in a spatial concept learning game. In *Proceedings of the 31st annual conference of the cognitive science society*, 3145–3150.

Gureckis, T.M. & Markant, D.B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, **7**, 464–481.

Harlé, K.M., Zhang, S., Schiff, M., Mackey, S., Paulus, M.P. & Yu, A.J. (2015). Altered statistical learning and decision-making in methamphetamine dependence: evidence from a two-armed bandit task. *Frontiers in psychology*, **6**, 1910.

Harlé, K.M., Guo, D., Zhang, S., Paulus, M.P. & Angela, J.Y. (2017). Anhedonia and anxiety underlying depressive symptomatology have distinct effects on reward-based decision-making. *PloS one*, **12**, e0186473.

Hills, T.T., Todd, P.M. & Goldstone, R.L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, **19**, 802–808.

Hills, T.T., Todd, P.M., Lazer, D., Redish, A.D., Couzin, I.D., Group, C.S.R. *et al.* (2015). Exploration versus exploitation in space, mind, and society. *Trends in cognitive sciences*, **19**, 46–54.

Itti, L. & Baldi, P.F. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, 547–554.

Jones, D.R., Schonlau, M. & Welch, W.J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, **13**, 455–492.

Kahneman, D. & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, **11**, 123–141.

Kalish, M.L., Lewandowsky, S. & Kruschke, J.K. (2004). Population of linear experts: knowledge partitioning and function learning. *Psychological Review*, **111**, 1072.

Kaplan, R., Schuck, N.W. & Doeller, C.F. (2017). The role of mental maps in decision-making. *Trends in neurosciences*, **40**, 256–259.

Kaufmann, E., Korda, N. & Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, 199–213, Springer.

Kemp, C. & Tenenbaum, J.B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, **105**, 10687–10692.

Kemp, C. & Tenenbaum, J.B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, **116**, 20.

Kidd, C. & Hayden, B.Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, **88**, 449–460.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A. *et al.* (2017).

Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, **114**, 3521–3526.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive science*, **18**, 513–549.

Kording, K.P. & Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. *Nature*, **427**, 244.

Kruschke, J.K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, **99**, 22.

Kuhn, D., Black, J., Keselman, A. & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, **18**, 495–523.

Lehman, J. & Stanley, K.O. (2011). Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, **19**, 189–223.

León-Villagrá, P., Preda, I. & Lucas, C.G. (2018). Data availability and function extrapolation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Lieder, F. & Griffiths, T.L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, **124**, 762.

Lieder, F., Plunkett, D., Hamrick, J.B., Russell, S.J., Hay, N. & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in Neural Information Processing Systems*, 2870–2878.

Lieder, F., Griffiths, T.L., Huys, Q.J. & Goodman, N.D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, **25**, 322–349.

Lloyd, K. & Leslie, D.S. (2013). Context-dependent decision-making: a simple bayesian model. *Journal of The Royal Society Interface*, **10**, 20130069.

Lucas, C.G., Griffiths, T.L., Williams, J.J. & Kalish, M.L. (2015). A rational model of function learning. *Psychonomic bulletin & review*, **22**, 1193–1215.

Maaten, L.v.d. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, **9**, 2579–2605.

Markant, D., Ruggeri, A., Gureckis, T.M. & Xu, F. (2016a). Enhanced memory as a common effect of active learning. *Paper in revision*.

Markant, D.B., Settles, B. & Gureckis, T.M. (2016b). Self-directed learning favors local, rather than global, uncertainty. *Cognitive science*, **40**, 100–120.

McCloskey, M. & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, vol. 24, 109–165, Elsevier.

McLeod, M., Osborne, M.A. & Roberts, S.J. (2018). Optimization, fast and slow: optimally switching between local and bayesian optimization. *arXiv preprint arXiv:1805.08610*.

Mehlhorn, K., Newell, B.R., Todd, P.M., Lee, M.D., Morgan, K., Braithwaite, V.A., Hausmann, D., Fiedler, K. & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G. *et al.* (2015). Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533.

Murdock, J., Allen, C. & DeDeo, S. (2017). Exploration and exploitation of victorian science in darwin's reading notebooks. *Cognition*, **159**, 117–126.

Murray, I. & Adams, R.P. (2010). Slice sampling covariance hyperparameters of latent gaussian models. In *Advances in Neural Information Processing Systems*, 1732–1740.

Nassar, M.R., Wilson, R.C., Heasly, B. & Gold, J.I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, **30**, 12366–12378.

Navarro, D.J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, **2**, 28–34.

Navarro, D.J., Griffiths, T.L., Steyvers, M. & Lee, M.D. (2006). Modeling individual differences using dirichlet processes. *Journal of mathematical Psychology*, **50**, 101–122.

Nelson, J.D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological review*, **112**.

Nilsson, H., Rieskamp, J. & Wagenmakers, E.J. (2011). Hierarchical bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, **55**, 84–93.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, **53**, 139–154.

Niv, Y. (2019). Learning task-state representations. *Nature neuroscience*, **22**, 1544–1553.

Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A. & Wilson, R.C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, **35**, 8145–8157.

Nosofsky, R.M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, **115**, 39.

O'Reilly, J.X. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in neuroscience*, **7**.

Parpart, P., Jones, M. & Love, B.C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, **102**, 127–144.

Qian, T., Jaeger, T.F. & Aslin, R.N. (2012). Learning to represent a multi-context environment: more than detecting changes. *Frontiers in Psychology*, **3**, 228.

Redish, A.D., Jensen, S., Johnson, A. & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, **114**, 784.

Renninger, L.W., Verghese, P. & Coughlan, J. (2007). Where to look next? eye movements reduce local uncertainty. *Journal of vision*, **7**, 6–6.

Reverdy, P.B., Srivastava, V. & Leonard, N.E. (2014). Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, **102**, 544–571.

Rothe, A., Lake, B.M. & Gureckis, T.M. (2016). Asking and evaluating natural language questions. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Ruggeri, A. & Lombrozo, T. (2014). Learning by asking: how children ask questions to achieve efficient search. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 1335–1340.

Sanborn, A.N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and cognition*, **112**, 98–101.

Sanborn, A.N., Griffiths, T.L. & Navarro, D.J. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society*, 726–731, Mahwah, NJ.

Sanborn, A.N., Griffiths, T.L. & Navarro, D.J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, **117**, 1144.

Schulz, E. & Gershman, S.J. (2019). The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, **55**, 7–14.

Schulz, E., Tenenbaum, J., Duvenaud, D.K., Speekenbrink, M. & Gershman, S.J. (2016). Probing the compositionality of intuitive functions. In *Advances In Neural Information Processing Systems*, 3729–3737.

Schulz, E., Klenske, E., Bramley, N. & Speekenbrink, M. (2017a). Strategic exploration in human adaptive control. *bioRxiv*, 110486.

Schulz, E., Konstantinidis, E. & Speekenbrink, M. (2017b). Putting bandits into context: How function learning supports decision making.

Schulz, E., Bhui, R., Love, B.C., Brier, B., Todd, M.T. & Gershman, S.J. (2018a). Exploration in the wild. *BioRxiv*, 492058.

Schulz, E., Wu, C.M., Ruggeri, A. & Meder, B. (2018b). Searching for rewards like a child means less generalization and more directed exploration. *bioRxiv*, 327593.

Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in cognitive sciences*, **16**, 382–389.

Schulz, L., Kushnir, T. & Gopnik, A. (2007). Learning from doing: Intervention and causal inference. *Causal learning: Psychology, philosophy, and computation*, 67–85.

Schulz, L.E. & Bonawitz, E.B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental psychology*, **43**, 1045.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**, 461–464.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P. & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, **104**, 148–175.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317–1323.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. *et al.* (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, **529**, 484–489.

Simon, H.A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, **69**, 99–118.

Sloman, S. & Lagnado, D. (2005). Do we 'do'? *Cognitive Science*, **29**, 5–39.

Snoek, J., Larochelle, H. & Adams, R.P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, 2951–2959.

Speekenbrink, M. & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, **7**, 351–367.

Speekenbrink, M. & Shanks, D.R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, **139**, 266.

Srinivas, N., Krause, A., Kakade, S.M. & Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Steyvers, M., Lee, M.D. & Wagenmakers, E.J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, **53**, 168–179.

Storn, R. & Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, **11**, 341–359.

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge.

Tassinari, H., Hudson, T.E. & Landy, M.S. (2006). Combining priors and noisy visual cues in a rapid pointing task. *Journal of Neuroscience*, **26**, 10154–10163.

Teodorescu, K. & Erev, I. (2014). On the decision to explore new alternatives: The coexistence of under-and over-exploration. *Journal of Behavioral Decision Making*, **27**, 109–123.

Thaker, P., Tenenbaum, J.B. & Gershman, S.J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, **77**, 10–20.

Vul, E., Goodman, N., Griffiths, T.L. & Tenenbaum, J.B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, **38**, 599–637.

Wang, J.X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D. & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Wang, Z. & Jegelka, S. (2017). Max-value entropy search for efficient bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3627–3635, JMLR. org.

Williams, C.K. & Rasmussen, C.E. (2006). Gaussian processes for machine learning. *the MIT Press*, **2**, 4.

Wilson, A.G., Dann, C., Lucas, C. & Xing, E.P. (2015). The human kernel. In *Advances in neural information processing systems*, 2854–2862.

Wilson, R. & Collins, A. (2019). Ten simple rules for the computational modeling of behavioral data.

Wilson, R.C. & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in human neuroscience*, **5**.

Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A. & Cohen, J.D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, **143**, 2074.

Wu, C.M., Schulz, E., Speekenbrink, M., Nelson, J.D. & Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. *bioRxiv*, 106286.

Wu, C.M., Schulz, E., Speekenbrink, M., Nelson, J.D. & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, **2**, 915.

Yi, M.S., Steyvers, M. & Lee, M. (2009). Modeling human performance in restless bandits with particle filters. *The Journal of Problem Solving*, **2**, 5.

Yu, A.J. & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, **46**, 681–692.